A Process Performance Monitoring Methodology for Mineral Processing Plants

by

Jacobus Willem De Villiers Groenewald

Dissertation presented for the Degree

of

DOCTOR OF PHILOSOPHY
(Extractive Metallurgical Engineering)

in the Faculty of Engineering
at Stellenbosch University

Supervisor
Prof. C. Aldrich

Co-Supervisors
Prof. S.M. Bradshaw
Prof. G. Akdogan

April 2014

Stellenbosch University http://scholar.sun.ac.za

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

JWD Groenewald 27/01/2014

…… ……..

Signature Date

Copyright © 2014 Stellenbosch University
All rights reserved

Stellenbosch University http://scholar.sun.ac.za

Abstract

Key to remaining competitive within the mineral industry is ensuring that all processes are always being operated optimally. Process performance monitoring is an ideal initiative with which to accomplish this. Not only can it be used to ensure fault free process operation, but it can also be applied for plant performance improvement through a better understanding of the contributors to the success or failure of the process operation. Critical to the success of any proposed monitoring approach would be its ability to cater for the fact that these mineral processes are typically highly complex, dynamic and non-linear.

The purpose of this study was to propose and evaluate a methodical approach to plant-wide process performance monitoring for mineral processing plants. Crucial to this approach is the concept of integrating process causality maps with data-based systems for event detection and diagnosis. To this end, process causality maps were developed to provide a means of structuring process data through the use of fundamental process knowledge. Statistical data-based fault detection techniques, being especially powerful with regards to data compression and dimensionality reduction, were employed to allow huge data sets to be analysed more easily. Change point detection techniques allowed for the identification of stationary segments of data in otherwise non-stationary data sets. Variable importance analysis was used to identify and interpret the variable(s) responsible for the event conditions.

Using simulated data sets, different techniques were evaluated in order to acquire an appreciation for their effectiveness and reliability. While it was found that no single technique significantly outperformed any other, it was confirmed that for data having different structures and characteristics, none of the techniques were effective in analysing all potential event conditions. It was suggested that all available techniques be run in parallel, with expert interpretation of the results, ensuring a more comprehensive analysis to be performed. Furthermore, given that only process measurements being monitored could be used to detect events and be analysed for importance, the consequences of monitoring too few process measurements were highlighted.

A generic analytical methodology for multivariate process performance monitoring was defined, ensuring the use of appropriate techniques and interpretations. The methodology was subsequently successfully applied to a mineral processing concentrator case study. The application of process causality maps was found to significantly simplify the challenge of monitoring the process, not only improving the ability of the techniques applied through a better focussed application, but also the interpretability of the results due to the reduction in complexity. Extreme learning machine, a robust and computationally inexpensive algorithm, was identified as a potential core algorithm for the data analysis techniques forming part of a process performance monitoring solution. With different drivers, at different times, having different effects on the process, visual representation of the data through canonical variate analysis biplots, combined with a sound understanding of the process under investigation, contributed significantly to a better understanding of the important variables for each event condition.

From an implementation perspective, adoption of the methodology remains the biggest barrier to success, requiring the most attention in the immediate future.

Stellenbosch University http://scholar.sun.ac.za

Opsomming

Krities tot mededingendheid in die minerale-industrie is die versekering dat alle prosesse altyd optimaal bedryf word. Proses prestasie-monitering is 'n ideale inisiatief waarmee dit bereik kan word. Nie net kan dit foutvrye proses werking verseker nie, maar dit kan ook toegepas word vir proses prestasie verbetering deur middel van 'n beter begrip van die bydraers tot die sukses of mislukking van die proses. Van kritieke belang vir die sukses van enige voorgestelde benadering tot monitering is die vermoë om voorsiening te maak vir die feit dat hierdie mineraal prosesse tipies hoogs kompleks, dinamies en nie-lineêr is.

Die doel van hierdie studie was om 'n metodiese benadering tot aanlegwye proses prestasie-monitering van mineraal aanlegte voor te stel en te evalueer. Deurslaggewend tot hierdie benadering is die integrasie van proses oorsaaklikheid kaarte met data-gebaseerde stelsels vir gebeurtenis opsporing en diagnose. Proses oorsaaklikheid kaarte is ontwikkel om voorsiening te maak vir die struktureering van proses data deur die gebruik van fundamentele kennis. Statistiese foutopsporings tegnieke, wat veral kragtig is ten opsigte van data kompressie en dimensionaliteit vermindering, is ingespan om groot datastelle makliker te ontleed. Veranderpunt opsporings tegnieke het toegelaat vir die identifisering van stasionêre segmente van data in andersins nie-stasionêre datastelle. Veranderlikebelang analise is gebruik om veranderlike(s) wat verantwoordelik is vir gebeurtenisse te identifiseer en te interpreteer.

Gesimuleerde datastelle is gebruik om verskillende tegnieke te evalueer ten einde 'n waardering vir hul doeltreffendheid en betroubaarheid te verkry. Alhoewel dit gevind is dat geen enkele tegniek aansienlik beter as enige ander is nie, is dit bevestig dat geen enkele tegnieke effektief in die ontleding van alle potensiële gebeurtenis was vir data met verskillende strukture en eienskappe. Daar is voorgestel dat alle beskikbare tegnieke in parallel uitgevoer word met deskundige interpretasie van die resultate om te versker data 'n meer omvattende analise uitgevoer word. Gegee dat slegs proses metings wat gemoniteer word gebruik kan word vir gebeurtenis opsporing en vir die belangrikheid ontleeding, is die nadeligheid van die monitering van te min proses metings uitgelig.

'n Generiese analitiese metodologie vir meerveranderlike proses prestasie-monitering is gevolglik omskryf om die gebruik van geskikte tegnieke en interpretasies te verseker. Hierdie metodologie is suksesvol op ‘n mineraal prosesserings aanleg gevallestudie toegepas. Dit is gevind dat die toepassing van die proses oorsaaklikheid kaarte die uitdaging van monitering aansienlik vereenvoudig, nie net as gevolg van die verbeterde vermoë van die tegnieke wat gebruik word deur 'n beter gefokusde toepassing nie, maar ook deur verberterde interpretering van resultate as gevolg van die vermindering in kompleksiteit. Uiterste leermasjiene, 'n robuust en bekostigbare berekeningsintensiewe algoritme, is geïdentifiseer as 'n potensiële kern algoritme vir die data-analise tegnieke, wat deel vorm van 'n proses prestasie-monitering oplossing. Visuele voorstellings het aansienlike bydraes gelewer tot beter begrip van belangrike veranderlikes vir elke toestand, veral deur middel van ‘n kombinasie van kanoniese veranderlike-onleding biplotte en 'n deeglike begrip van die proses wat ondersoek word.

Die aanvaarding van die metodologie is tans die grootste hindernis tot sukses vanuit 'n uitvoerings perspektief en benodig dus die meeste aandag in die onmiddellike toekoms.

Stellenbosch University http://scholar.sun.ac.za

Acknowledgements

I would like to take this opportunity to thank all those that united against me to make this dissertation a reality – it has been the most momentous academic challenge that I have ever faced. In particular, the following people deserve a special mention:

First and foremost, I would like to convey a sincere word of gratitude to Anglo American Platinum for sponsoring this project.

I would like to thank Professor Chris Aldrich, not only for his technical and theoretical assistance, but also his guidance and support.

I would also like to thank my parents for giving me the opportunity to live, to learn, and to love.

Finally, I would like to dedicate this thesis to my loving wife, Annette. Thank you for all your patience, support and tireless encouragement.

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom."

- Clifford Stoll & Gary Schubert

Stellenbosch University http://scholar.sun.ac.za

Table of contents

Stellenbosch University http://scholar.sun.ac.za

Stellenbosch University http://scholar.sun.ac.za

Stellenbosch University http://scholar.sun.ac.za

Stellenbosch University http://scholar.sun.ac.za

Stellenbosch University http://scholar.sun.ac.za

Abbreviations

AANNPCA	Auto-associative neural network principal component analysis
ACP	Anglo American Platinum Converting Process
AF	Analysis Framework
AMI	average mutual information
ANOVA	analysis of variance
APC	advanced process control
ARMA	autoregressive-moving-average
BRPM	Bafokeng-Rasimone Platinum Mine
CART	Classification and Regression Trees
CELM	ELM-with-Bagging
Cur	current
CUSUM	cumulative sum
Conc	concentrate
CR_{MPP}	cleaning/regrinding flotation stage
CT	classification tree
CVA	canonical variate analysis
d50c	distribution 50% cut-point
DAPCA	delay-adjusted principal component analysis
DISSIM	process data dissimilarity index
DPCA	dynamic principal component analysis
ELM	extreme learning machine
EWMA	exponentially weighted moving average
FC	flow controller
FC_{MPP}	flotation circuit
FDA	Fisher discriminant analysis
FFloat	chrome circuit tails (flash flotation tails)
FI	flow indicator
FIC	flow indicator controller
FT	flotation cell
FTail	final tails
GC_{MPP}	grinding circuit
GTST	goal tree-success tree
GUI	graphical user interface
iid	independent and identically distributed
JV	joint venture
KBS	knowledge-based system
KPI	key performance indicator
kPa	kilopascal

Stellenbosch University http://scholar.sun.ac.za

kW	kilowatt
kWh/t	kilowatt hour per ton
LDA	linear discriminant analysis
LIC	level indicator controller
m^3/hr	cubic meter per hour
mm	millimetre
MLP	multilayer perceptron
MPCA	moving principal component analysis
MSE	mean square error
MSPCA	multiscale principal component analysis
MTS	multivariate time series
MV	manipulated variable
ND-PCA	non-linear dynamic principal component analysis
NIPALS	non-iterative partial least squares
NLMSPCA	non-linear multiscale principal component analysis
NLPCA	non-linear principal component analysis
NNPCA	neural network principal component analysis
NOC	normal operating condition
OOB	out-of-bag
OOCC	out-of-control condition
OPM	operational performance monitoring
PC	principal component
PCA	principal component analysis
PGM	platinum group metal
PI	pressure indicator
PIC	pressure indicator controller
PLS	partial least squares
PMR	Precious Metals Refinery
PSD	particle size distribution
RBMR	Rustenburg Base Metals Refiners
Ref	reference
RF	random forests
ROC	rate of change
RS_{MPP}	rougher/scavenger flotation stage
RTail	primary (rougher) flotation tails
SC	scatter matrix
SLFN	single hidden layer feedforward neural network
SMP	simple multivariate process
SP	set point
SPC	statistical process control

Stellenbosch University http://scholar.sun.ac.za

SPE	squared prediction error
SPM	statistical process monitoring
SSA	singular spectrum analysis
SSPCA	summed-scores principal component analysis
SVD	singular value decomposition
t/m^3	tons per cubic meter
TEP	Tennessee Eastman process
TI	temperature indicator
TIC	temperature indicator controller
tph	tons per hour
UCL	upper confidence level
µm	micrometre
VIF	variance inflation factor
VM	virtual machine
WLTR	Western Limb Tailings Retreatment

Stellenbosch University http://scholar.sun.ac.za

Nomenclature

A	moving PCA statistic for monitoring changes in direction
A_{1-m}	moving PCA statistic for monitoring changes of subspace
$A(j)$	principal curves accumulated scores
A_{TE}	Tennessee Eastman reactant
A_{SMP}	Simple multivariate process constant
a	number of PCs retained / number of classification trees / model accuracy
a_m	vector of wavelet scaling function coefficients
B	cost complexity measure / overall misclassification rate
B_{TE}	Tennessee Eastman inert
B_{SMP}	Simple multivariate process constant
b	coefficient
C	change condition
C_{TE}	Tennessee Eastman reactant
C_{SMP}	Simple multivariate process constant
c	variable / intermediate vector
c_f	projection index
D	dissimilarity index
D_n	sum of squared Euclidean distances
D_{TE}	Tennessee Eastman reactant
D_{SMP}	Simple multivariate process constant
d	process delay
d_m	vector of wavelet coefficients
E	residual/noise matrix
E	expectation operator
E_{TE}	Tennessee Eastman reactant
e	residual/noise vector
F	distribution
$\Gamma(\bullet)$	non-linear principal curve function
F_{TE}	Tennessee Eastman byproduct
f	principal curve

Stellenbosch University http://scholar.sun.ac.za

G	residual matrix
G_{TE}	Tennessee Eastman product
g	weight of weighted chi-squared distribution
$g(x)$	activation function
H	distribution / hypothesis
H_{TE}	Tennessee Eastman product
h	degrees of freedom of weighted chi-squared distribution
I	filter
i	counter
J	filter
j	time step / observation number
K	embedding dimension
k	class labels
L	loading matrix
l	eigenvectors/loadings/principal components / Fisher discriminant analysis vector
M	lag parameter
m	number of principal components / variables / wavelet scale / optimal random split parameter
n	number of rows/observations
N	number of rows/observations / window width
$\tilde{N}$	Hidden node number
O	number of subsets
P	orthogonal projection / transformation matrix / hidden layer output matrix
$P(\ )$	probability
p	number of columns/variables/dimensions
$\hat{p}$	fraction of samples
Q	squared prediction error lack of model fit statistic / decision tree node impurity measure
q	eigenvectors/loadings/principal components
R	covariance matrix / decision tree logical region
r	number of linear relations / run length / rank
S	state variable
S	cumulative sums index
S_{diff}	cumulative sums difference
S_n	normalised sum of squares of distances
s	dilation parameter

Stellenbosch University http://scholar.sun.ac.za

T	distribution
T^2	Hotelling's statistic
T	non-linear principal score matrix
t	scores/latent vectors/variables
u	scores/latent vectors
v	translation parameter
W	moving window data set / CUSUM-type statistic
w	window size / time lags / time window / test sample size
X	input data matrix
x	input vector
Y	output data matrix
y	output vector
Z	weight matrix
z	weight vector
α	significance level
β	output weights
χ	hyperparameter
χ^2	weighted chi-squared distribution
δ	number of classes
ε	random error
ϕ	change point
φ	tree
η	tree node
κ	class outcome
λ	eigenvalue / estimate of variance
μ	mean
ν	hyperparameter
π	predictive probability
σ	standard deviation
τ	time window
υ	variance
ω	measure of variable importance
ψ	wavelet
ζ	complexity parameter
Λ	diagonal matrix
Θ	probability parameter

Stellenbosch University http://scholar.sun.ac.za

1 Introduction

One of the key factors for the mineral industry to remain competitive in today's economy is for their various extraction and metallurgical processes to always be operating optimally. In order to know whether or not a process is operating optimally, process performance information and its associated information management is crucial. This process performance information in turn allows personnel to not only detect, identify and correct faulty process operational parameters, control and/or equipment, but also allows for the improvement of plant performance through initiatives such as asset optimisation and continuous process improvement.

Many mineral processing plants these days, especially in the platinum industry, record numerous high frequency operational data. This allows for the effective monitoring of process performance with the aim of ensuring fault free process operation. These mineral processes that require monitoring are typically highly dynamic and non-linear, with data overload commonly occurring. Not only is the automated monitoring of process key performance indicators (KPIs) sought after, but ad hoc analysis is included as part of the process performance monitoring for when the detection and associated root causes of abnormal plant conditions are not that obvious. It is becoming increasingly necessary to understand the causes of reduced performance and the relationship between the KPIs in order to extract the greatest economic benefits.

Over the last few years, the effective monitoring of critical process KPIs within mineral processing plants has become a reality. This is a direct result not only from improvements made in measurements sensors and computer systems, but also due to advances made in the techniques used to analyse the data. Modern equipment not only include on-stream analysers, measuring previously unavailable concentrate flows, grades and particle size distributions, but also allows for the storage and management of large process data sets resulting from these measurement sensors. Unfortunately, the majority of process performance monitoring applied in the mineral processing industry is often through the haphazard, informal inspection of trends and tables representing various process data. The aim of this study is therefore the introduction of a formal process performance monitoring methodology, providing a structured approach and relevant data analysis techniques with which to assist plant personnel during the task of fault detection, diagnosis and correction. This should not only lead to earlier detection of abnormal process conditions, but also to more effective fault diagnosis and corrective action.

1.1 Monitoring objectives

The primary objective of process performance monitoring in the context of this study is to ensure fault free process operation through the combined monitoring of process KPIs and subsequent detection, identification and correction of faulty process operational parameters, control and/or equipment. As a secondary objective, process performance monitoring is applied for the improvement of plant performance through a better understanding of the contributors to the success or failure of a plant as well as initiatives such as asset optimisation and continuous process improvement.

Stellenbosch University http://scholar.sun.ac.za

Breiman (2001) defines two typical goals in analysing data: prediction (predicting the response to future input variables) and information (extracting information about how input variables are associated to response variables). With these goals in mind the process performance monitoring objectives have been translated to the Anglo American Platinum business objectives of providing expert support to operations by:

- **Proactively detecting shifts in process performance:** *Using process knowledge and models to identify shifts in performance and provide a structured means of determining the root cause of the shifts and, if required, propose corrective action.*
- **Identifying process improvement opportunities:** *Provide sufficient information and root cause diagnostics to identify areas of improvement and new performance indicators.*

Successfully achieving these business objectives necessitates an integrated data analysis system that runs with little or no human intervention – in part requiring the integration of the various data sources present on the process operations, providing a single consistent source of information for the operations technical staff. Furthermore, the data analysis techniques, when applied to continuous time series data, needs to be applicable to static (steady state) and dynamic, linear and non-linear, as well as univariate and multivariate systems. Other critical considerations to take into account include:

- Ability to ***detect "changes"*** in the data being monitored. The expected "changes" to be detectable should be clearly defined and, consequently, will place a limitation on the usefulness of the methodology.
- ***Robustness***, ***sensitivity to "changes"***, and ***detection speed*** are critical. Predictive monitoring could be used to greatly improve detection speed.
- If used for event diagnosis, ***accurate event diagnosis*** is critical.

1.1.1 Anglo American Platinum

Anglo American Platinum is a member of the Anglo American plc group. Anglo American Platinum's mining, concentrating, smelting and refining operations are situated in the North West, Limpopo and Mpumalanga provinces of South Africa (Figure 1) as well as Zimbabwe. The company produces platinum group metals (PGMs) in quantities determined by their occurrence in the ore mined. The PGMs are platinum, palladium, rhodium, ruthenium, iridium and osmium. Nickel, copper, gold and small quantities of other base metals are by-products of PGM operations.

Stellenbosch University http://scholar.sun.ac.za

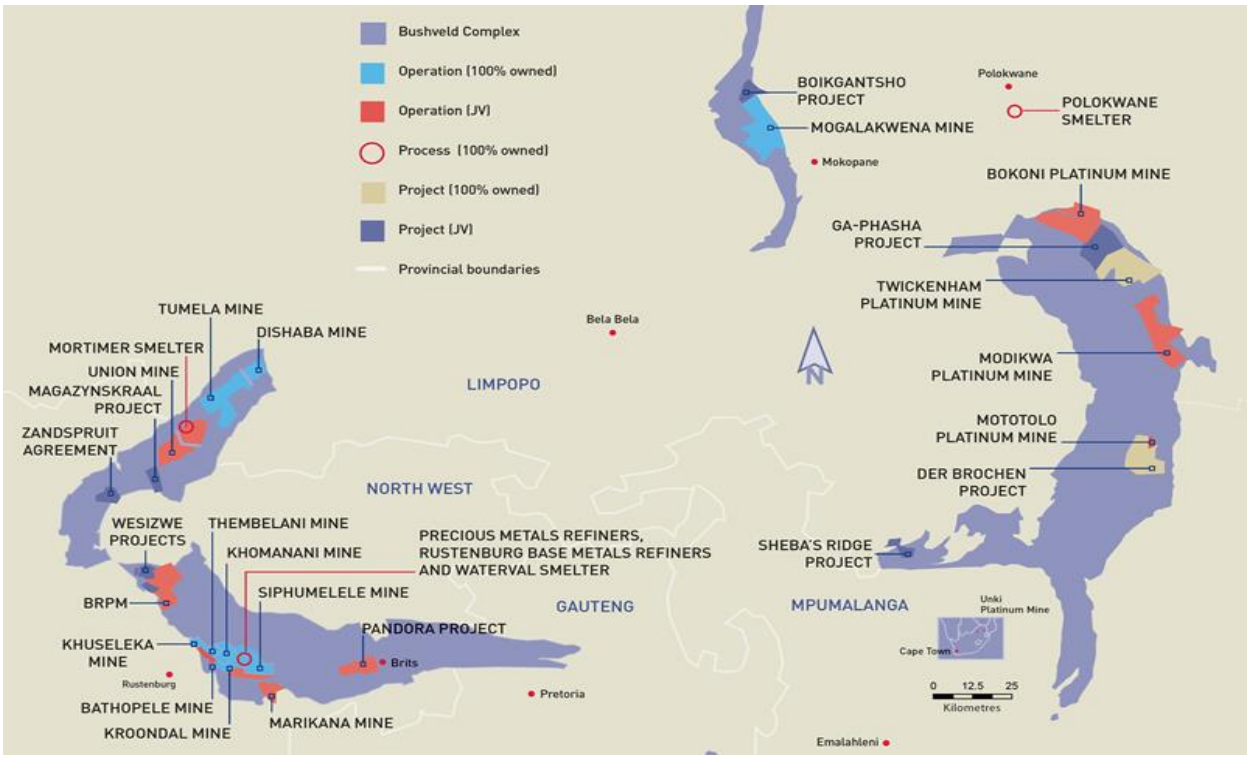

Figure 1: **Geographical overview of Anglo American Platinum process operations (Bushveld complex, 2011)**

Anglo American Platinum's mineral processing operations consist of concentrator plants, smelters and refineries (Figure 2). The concentrator plants treat ore from the mining operations to produce PGMs and base metals in the form of concentrate. This product is further processed at the three smelters (Mortimer, Polokwane and Waterval) and two refineries (Rustenburg Base Metals Refiners (RBMR) and Precious Metals Refinery (PMR)) to produce mainly saleable metals.

Each of the mineral processing operation sites can be sub-divided into process areas, process cells, process units and equipment modules. This allows data analysis techniques to not only be applied to each mineral processing operation site as a whole (plant-wide), but also to individual process areas, process cells, process units and equipment modules. Data can thus be better structured for data analysis with the aim of meeting the business objectives, potentially also improving speed and accuracy during analyses. As is often the case, various philosophies exists whereby the data can be structured, causality captured and the analyses performed.

Stellenbosch University http://scholar.sun.ac.za

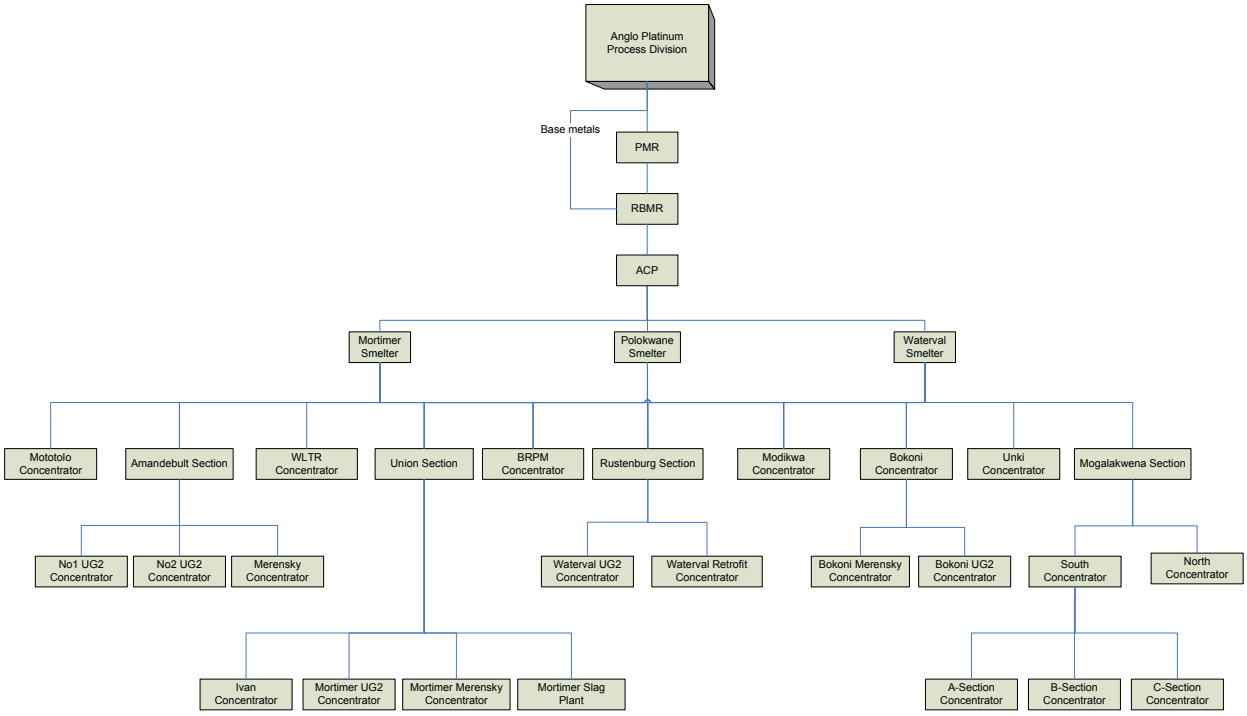

Figure 2: **Hierarchical overview of Anglo American Platinum process operations**

1.2 Process performance monitoring

Process performance monitoring, as opposed to process fault monitoring, not only aims to monitor the process for fault conditions but also for opportunities of improvements. Core to the success of either of these monitoring strategies are the techniques on which they rely. Depending on the type of *a priori* knowledge used, the techniques can be classified as either fundamental model-based (Blanke et al., 2006; Ding, 2008; Gertler, 1998; Patton, 2000) or process data-based (Ding et al., 2011; Qin, 2003; Qin, 2009; Thornhill and Horch, 2007; Venkatasubramanian et al., 2003c). If the *a priori* domain knowledge is developed from a fundamental understanding of the process using first-principles knowledge, it is referred to as model-based knowledge. If the *a priori* domain knowledge is gleaned from past experience with the process, it is referred to as process data-based knowledge. Due to the fact that fundamental model-based approaches are often very complex and costly to develop, the simpler process data-based approaches have become more popular for large-scale industry applications (Chiang et al., 2001; Russell et al., 2000). With several studies having reported the potential of and benefits derived from the use of fundamental process knowledge with data-base models (Van Lith et al., 2003; Venkatasubramanian et al., 2003c; Nandi et al., 2004; Van Sprang et al., 2005; Thornhill and Horch, 2007; Yang and Xiao, 2012), the focus of this study is also on the use of process data-based approaches, derived from historical process data, supported by limited fundamental model-based approaches.

Stellenbosch University http://scholar.sun.ac.za

Academically the mineral processing industry has received some attention with regards to the development of new, and improvement of existing, data-based models and statistical data analysis techniques related to process performance monitoring (Hodouin et al., 1993; MacGregor and Kourti, 1995; Jemwa and Aldrich, 2006). These models and techniques have, however, historically been applied in isolation, not forming part of an overarching methodology for process performance monitoring.

Practically the majority of process performance monitoring applied in the mineral processing industry is through the inspection of trends and tables of raw instrument data and calculated process KPIs with very little, if any at all, information added to the data using statistical data analysis to either highlight areas of importance or reduce the number of data looked at. Furthermore, for areas of importance where something went wrong in the process it is also not always good enough to only identify when something went wrong and what the root cause of the problem is. Often an understanding of how the bottom line of the company is affected is needed, highlighting the importance of fault mitigation.

1.2.1 Process causality maps

Although work has been done on process performance monitoring in the mineral processing industry, relatively little attention has been given to plant-wide systems. Whereas representing the entire plant as a single large matrix of variables is theoretically possible, producing a convenient representation of the process with all variables being considered simultaneously, reliably detecting fault conditions become extremely difficult due to the possibility of systematic errors getting lost among all the variables. Such a representation also does not allow for the exploitation of existing knowledge with regards to the relationships between variables e.g. upstream and downstream, recirculation, etc.

Whereas the use of quantitative model-based approaches in real industrial processes is often limited due to the difficulty in developing accurate mathematical system models based on complex, highly dimensional, non-linear processes, qualitative model-based approaches such as causal models and abstraction hierarchies are a viable alternative. Qualitative model-based approaches not only capture the causal structure of a system in a more insightful manner than conventional expert systems but are also not as rigid in nature as numeric simulations. This not only allows for the representation of the system but also the ability to reason about it. Furthermore, it has been shown that hierarchical decompositions of the processing systems provide effective modularity for organising large-scale diagnostic knowledge-bases as well as allowing different techniques to be integrated to address specific local problems (Prasad et al., 1998). It should, however, be noted that although focussing on individual units in isolation may improve detection, identification of actual fault conditions may become more difficult.

To this end, ***process causality maps*** have been developed to be used as a framework for plant-wide root cause analysis as part of the process performance monitoring methodology. When considering the trade-off between complexity and gain when compared to monitoring an entire process as a single system (all variables simultaneously), process causality maps are a much simpler and more structured approach allowing for both the monitoring of independent process units (using subsets of variables) and the relationships between related process units.

Stellenbosch University http://scholar.sun.ac.za

1.2.2 Statistical data-based fault detection

From a process data-based perspective qualitative approaches such as expert systems suffer limitation in that they need to be developed too specifically, are limited in terms of representation power and are often very difficult to update (Rich and Venkatasubramanian, 1987), making them less ideal when compared to quantitative approaches for use in process performance monitoring. Quantitative process data-based methods essentially formulate the diagnostic (detect and identify) problem-solving as a pattern recognition problem, with statistical methods using knowledge of *a priori* class distributions. Multivariate statistical techniques are especially powerful with regards to data compression, dimensionality reduction and the handling of noise and correlation ensuring the retention of essential process information, thus allowing huge data sets to be more easily analysed (Venkatasubramanian et al., 2003c). To this end, statistical process control (SPC) makes a valuable addition to process causality maps in the context of ***statistical data-based fault detection*** for process performance monitoring.

1.2.3 Change point detection

Fault detection is, however, only one component of process performance monitoring. Fault detection techniques are typically specifically applied for monitoring process deviations, focussing mainly on comparing new process data to a set of fixed, normal operating process data (where the reference/normal operating process data remain constant), thus detecting any deviation from normal process behaviour. Abnormal process operation is consequently inferred when a monitoring statistic exceeds a predetermined warning or control threshold. Such an approach is limited in its functional use in that it ignores potential events, wanted or unwanted change points, which can occur while the process is in a state of either normal or abnormal process operation (where the reference/normal operating process data is continually updated). Change points in data can be viewed as times of discontinuities or changes in process behaviour induced for example from changes in a process, input conditions, equipment and/or measurement techniques. These as yet undetected change points can potentially be used to identify process improvement opportunities and even contribute considerable knowledge about the process operation. Change point detection, quantitative process data-based methods used for estimating the time of change in a process, are complementary to fault detection, although not as well established in process applications. Contrary to statistical data-based fault detection, change point detection aims to determine whether or not a process in its current state is exhibiting a change in behaviour compared to the behaviour it exhibited immediately preceding its current state, irrespective of whether it was in a state of normal process behaviour or not. Change point detection extends the capability of process performance monitoring through its ability to solve problems that fault detection cannot: detection and analysis of potential interesting events that are occurring while the process is in a state of either normal or abnormal process operation. Change point detection effectively identifies stationary segments of data in an otherwise non-stationary data set with the stationary segments being of various lengths, magnitudes, or even patterns. To this end, ***change point detection*** makes a valuable addition to process performance monitoring in the context of event detection.

Stellenbosch University http://scholar.sun.ac.za

1.2.4 Variable importance analysis

Once a process change has been detected, either through fault detection or change point detection, the focus typically shifts to identifying the variable(s) responsible for the process change. For fault detection, contribution plots are predominantly used in identifying the relevant variables responsible for process fault conditions. Situations do however arise when contribution plots are not suitable, possibly due to the underlying fault detection model being inaccurate, or not reliable enough for use. Furthermore, for potentially interesting process events detected by change point detection, where the application of fault detection models is unsuitable or unreliable, an alternative to contribution plots are also needed. Variable importance analysis, which traditionally has been associated with the diagnosis of complex models, is a viable alternative to contribution plots. When modelling process event conditions, variable importance analysis can easily be used to identify the variable(s) responsible for the process event conditions. As with change point detection being complementary to fault detection, variable importance analysis can be seen as being complementary to SPC variable contribution plots. To this end, ***variable importance analysis*** makes a valuable addition to process performance monitoring in the context of event diagnosis.

1.3 Thesis objective

The main objective of this study is to propose and evaluate a methodical approach to plant-wide process performance monitoring for mineral processing plants. The approach is based on the concept of integrating process causality maps, providing a means of structuring process data through the use of fundamental process knowledge, with data-based systems, for event detection and diagnosis. This objective is achieved by attaining the following goals:

- The definition of process causality maps through the integration of causality with hierarchical system decomposition.
- A critical literature survey on data-based statistical process monitoring techniques for fault detection.
    - The development and implementation of a novel extreme learning machine (ELM) based fault detection technique: ELM principal component analysis.
    - A comparative evaluation of the various techniques on simulated time series data, simulated process data and a benchmark chemical plant simulation.
- A review of selected data-based change point detection techniques for inclusion in the process performance monitoring methodology in the context of event detection.
    - The development and implementation of a novel ELM-based change point detection technique.
    - The testing and comparison of the techniques for the detection of process event conditions in simulated time series data, simulated process data and a benchmark chemical plant simulation.
- A review of selected data-based variable importance analysis techniques for inclusion in the process performance monitoring methodology in the context of event diagnosis.

Stellenbosch University http://scholar.sun.ac.za

  - The development and implementation of a novel ELM-based variable importance analysis technique.
  - The evaluation and performance comparison of the techniques with regards to the identification and interpretation of important variables related to process event conditions in simulated time series data, simulated process data and a benchmark chemical plant simulation.
- The development of an analytical methodology to plant-wide process performance monitoring for mineral processing plants, integrating process causality maps and data-based methods.
  - The development of process causality maps for a PGM concentrator plant through the integration of causality with hierarchical system decomposition.
  - The evaluation of this methodology through a comprehensive real-world mineral processing case study.
  - A critical evaluation of the practical issues with regards to the implementation of the methodology.

1.4 Thesis layout

Chapter 2 presents the background to plant-wide process performance monitoring and the concept of process causality maps. Chapter 3 presents an overview and critical evaluation of statistical data-based fault detection techniques. Chapter 4 presents an overview and critical evaluation of a select few change point detection techniques, with chapter 5 similarly presenting an overview and critical evaluation of a select few variable importance analysis techniques. Chapter 6 defines an analytical methodology for process performance monitoring together with the derivation of process causality maps for a mineral processing concentrator process and its operating units. Chapter 7 is a comprehensive mineral processing industrial case study based on the proposed process performance monitoring methodology. Chapter 8 reviews some of the practical issues related to the implementation of a process performance monitoring methodology. Lastly, chapter 9 closes with conclusions and recommendations.

Stellenbosch University http://scholar.sun.ac.za

2 Plant-wide process performance monitoring

Process performance monitoring, in the context of this study, is aimed at ensuring fault free process operation through the combined monitoring of process KPIs and the subsequent detection, identification and correction of faulty process operational parameters, control and/or equipment. Central to this is the concept of fault detection of diagnosis. Whereas fault detection deals with the timely detection of abnormal process events, fault diagnosis deals with determining the causal origins of any detected abnormal process event.

Practically, the majority of process performance monitoring applied in the mineral processing industry is through the manual, *ad hoc* inspection of trends and tables of raw instrument data and calculated process KPIs. Relying on such an unstructured approach to process performance monitoring is not only made difficult due to its dependence on suitably knowledgeable metallurgical staff at different processing operations, but also due to information overload resulting from the size and complexity of the processing operations, the variety of potential process events needing detection and diagnosis, and the potentially unreliable nature of the process measurement data being analysed. Venkatasubramanian et al. (2003a) highlighted the importance of automating fault detection and diagnosis, providing assistance to operational staff in detecting and diagnosing process events. Core to such automation is the definition of a structured approach, a methodology, and selection of appropriate techniques for process performance monitoring.

Process performance monitoring techniques can be classified based on the type of *a priori* knowledge used in deriving the techniques (Venkatasubramanian et al., 2003a). If the *a priori* domain knowledge used stems from a fundamental understanding of the process using first-principles knowledge, the technique can be classified as fundamental model-based (Blanke et al., 2006; Ding, 2008; Gertler, 1998; Patton, 2000). If the *a priori* domain knowledge used stems from past experience with the process, the technique can be classified as process data-based (Ding et al., 2011; Qin, 2003; Qin, 2009; Thornhill and Horch, 2007; Venkatasubramanian et al., 2003c).

For model-based approaches, developing accurate mathematical models can be particularly difficult. Various factors contribute to the challenge of modelling a system, the most prominent being process non-linearity, high dimensionality, lack of good data and system complexity. Due to the complexity and costly nature of developing model-based approaches, their usefulness in real industrial processes is limited (Venkatasubramanian et al., 2003a). Contrary to model-based approaches, data-based approaches require very little modelling effort, making them relatively easy to implement and more popular for large-scale industry applications (Chiang et al., 2001; Russell et al., 2000). With fault detection traditionally being of greater than fault diagnosis, statistical data-based approaches have historically been well studied and successfully employed in industrial applications (Venkatasubramanian et al., 2003c).

This chapter gives an overview of plant-wide process performance monitoring approaches and discuss some of the theory behind process causality maps and its potential application to process performance monitoring.

Stellenbosch University http://scholar.sun.ac.za

2.1 Statistical data-based fault detection

Statistical data-based fault detection formulates the challenge of fault detection and diagnosis as a pattern recognition problem. This is accomplished through the use of *a priori* class distributions to classify data into pre-determined classes. When a process can be considered to be within a "state of statistical control", while still allowing "common-cause" variation, the probability distributions of the resulting process data should correspond to those of the process while in a state of normal operating condition (NOC). When the process can be deemed to be out of control, a change will occur in the underlying probability distributions of the process data, which, when using parametric techniques, will manifest itself as changes in parameters characterising the probability distributions. Statistical models are normally built using historical process data from when the process was in a state of NOC. New process data is then compared to this NOC statistical model in order to detect changes in the system.

Conventional univariate quality control charts, such as the Shewhart chart (Shewhart, 1931), the cumulative sum (CUSUM) chart (Page, 1954) and the exponentially weighted moving average (EWMA) chart (Roberts, 1959), are some of the earliest approaches to statistical data-based fault detection. These charts are based on the assumption that a process will remain within a "state of statistical control", while still allowing "common-cause" variation, with some process variables remaining close to their desired values. Consequently, conventional univariate control charts all monitor the deviation, or variation thereof, between a single process variable and its associated target value, independent of all other variables. When presented with correlated variables, the use of univariate control charts may be misleading, often resulting in inaccurate, delayed conclusions being drawn (Kourti and MacGregor, 1995).

In contrast to univariate control charts, multivariate statistical techniques extract information on how multiple process variables behave relative to each another while simultaneously compressing the data, reducing its dimensionality, in a manner which retains essential information. Principal component analysis (PCA) and partial least squares (PLS) are standard multivariate statistical projection and regression methods, able to handle high dimensional and correlated variables while extracting true information effectively (Wise and Gallagher, 1996). Whereas conventional PCA is used to find linear combinations of the process variables in a data set, effectively reducing the number of correlated process variables to a smaller set of uncorrelated factors, PLS attempts to find factors that not only capture the greatest amount of variance in the process variables but also the variation in the process variables most predictive of the product variables.

Various statistics exist with which to monitor/validate PCA and PLS models. The most common of these is the Hotelling's T^2 statistic (Hotelling, 1947) and the lack of model fit statistic, Q (Kourti and MacGregor, 1995). The Hotelling's T^2-chart provides a test for deviations in the variables that are of greatest importance to the variance of the data set, being able to detect only whether the variation in the variables in the plane of the selected principal components are greater than can be explained by "common-cause". The Squared Prediction Error (SPE) chart, based on the Q statistic, measures the amount of variation in each observation not captured by the selected principal components retained in the

Stellenbosch University http://scholar.sun.ac.za

model, indicating abnormalities resulting from when the relationship among variables or their dynamic patterns change. A third, fairly common, multivariate control chart is the score-chart, constructed by plotting the scores obtained from applying PCA to the data set in either two or three dimensions. The score-chart is used to indicate abnormalities resulting from several variables having a larger than normal change, whereas the basic relationship among the variables and their dynamic patterns do not change.

Kano et al. (2001) improved upon conventional PCA by applying PCA to a moving time-window data set, called moving PCA (MPCA), and deriving suitable monitoring indices. In this method, changes in the NOC, characterised by changes in the correlation structure among the process variables and not changes in the scores of predefined principal components, are detected through changes in the direction of each principal component or changes in the subspace spanned by several principal components. Kano et al. (2002) also introduced the process data dissimilarity index which, in contrast to the Hotelling's T^2 and the Q statistics, is based on the idea that a change in operating condition can be detected by monitoring the distribution of the process data reflecting the corresponding operating condition. More recently, local and global PCA (Yu, 2012), a manifold learning algorithm, and a new version of weighted PCA (Jiang and Yan, 2012) have been proposed. Local and global PCA not only preserve the global variance information of the Euclidean space, but also preserves the local structure, capturing more meaningful process information and thus improving upon the data separation capability of conventional PCA. Weighted PCA makes use of the change rate of the T^2 statistic along every principal component in order to capture the most useful information in the process, setting different weighting values for the selected principal components to highlight useful information when applied for online monitoring.

Although conventional PCA caters for the multivariate nature of process data, it does not consider the potential dynamic behaviour of the process. With the current values of any variable depending on the past values of that variable for dynamic systems, Ku et al. (1995) extended conventional PCA through a well-known time lagged variable method to include dynamic behaviour in the PCA model, called dynamic PCA (DPCA). DPCA has subsequently been extended to cater for possible non-linear relationships among process variables (Lin et al., 2000) as well as the modelling of known fault conditions (Tsung, 2000). In contrast to DPCA, Kruger et al. (2004) proposed the application of autoregressive-moving-average (ARMA) filters to remove auto-correlation from PCA derived score variables, reducing the number of false alarms potentially produced by strongly auto-correlated score variables, while Detroja et al (2007) used correspondence analysis for an improved representation of the dynamic correlation in process data. Wachs and Lewin (1999) catered for the dynamic behaviour of processes by either recursively summing the PCA scores, referred to as summed-scores PCA (SSPCA) or by applying relative shifts between the input and output variables, referred to as delay-adjusted PCA (DAPCA). In order to model data containing contributions from events whose behaviour changes over both time and frequency, Bakshi (1998) suggested the use of multiscale PCA (MSPCA), combining the ability of PCA for linearly decorrelating variables with that of wavelet analysis for decorrelating auto-correlated measurements. MSPCA has subsequently been extended to cater for possible non-linear relationships among process variables (Fourie and de Vaal, 2000).

Stellenbosch University http://scholar.sun.ac.za

Although great success has been achieved with the application of linear dimensionality reduction techniques, non-linear relationships exist in many real processes, requiring non-linear extensions to statistical data-based fault detection techniques. Various different approaches have been proposed with which to cater for non-linear characteristics in data, some of which include combining PCA and neural networks (Jia et al., 1998; Chen and Liao, 2002), combining principal curves and neural networks (Dong and McAvoy, 1996; Zhang et al., 1997) and using auto-associative neural network for deriving non-linear principal components (Kramer, 1991; Scholz et al., 2008). With the potential shown by kernel methods for non-linear fault detection and diagnosis (Lee et al., 2004; Choi et al., 2005; Cho et al., 2005), Jemwa and Aldrich (2006) proposed the use of kernel PCA for feature extraction combined with one-class support vector machines with which to estimate nonparametric confidence limits for the resulting monitoring charts. Deng and Tian (2013) subsequently combined kernel PCA, for non-linear transformation, and statistics pattern analysis, for statistics information mining, for process fault pattern recognition, with Qingchao and Xuefeng (2013) developing a fault detection method based on sensitive kernel principal component analysis where the most sensitive kernel principal components, based on the T^2 statistic, is used for fault detection.

2.2 Expert systems & trend analysis

Although not as widely studied for fault detection as statistical data-based approaches, expert systems and trend analysis have been shown to be viable data-based approaches to the challenge of fault detection and diagnosis. Expert systems imitate human expert knowledge in solving problems by providing a platform for capturing the reasoning processes of experts as rules. It is expected that such systems will outperform their human counterparts in making effective decisions pertaining to specific areas of expertise, especially being more consistent in their reasoning for a particular set of process conditions. Given the transparent nature of reasoning within an expert system, the ability of expert systems to provide an explanation for the provided solution is particularly powerful.

Simply stated, an expert system for fault detection and diagnosis consist of if-then rules that are used to organise expert knowledge regarding fault conditions (Kumamoto et al., 1984). Due to the transparent nature of experts systems, it is possible not only to identify a fault condition and its cause, but also the exact rules and reasons as to why it was identified. Expert systems for fault detection and diagnosis have also been extended through the use of hierarchical classification for structuring the knowledge-base (Ramesh et al., 1988), using an object-based knowledge representation for model and heuristic-based knowledge (Basila et al., 1990), neural networks as a first-level filter for diagnosing commonly encountered faults within chemical process plants (Becraft and Lee, 1993) and signed directed graphs as a process model with a fuzzy logical evaluation of the if-then rule base (Tarifa and Scenna, 1997).

Qian et al. (2003) developed an expert system for real-time fault diagnosis where the knowledge representation is divided into facts, functions and production rules. Based on the time varying nature of processes, the facts are sub-divided into static or dynamic facts, the production rules are sub-divided into general and time constraint extended production rules, and the functions employed to improve the numerical computational ability of the system. Contrary to this, Nan et al. (2008) proposed the use of

Stellenbosch University http://scholar.sun.ac.za

process trend recognition and fuzzy logic in a real-time fault diagnosis expert system making. Whereas process trend recognition is used to establish the relationship between the process operations and the sensor trends, fuzzy logic is used to map the expert knowledge with process trends using if-then rules. Real-time fault diagnosis expert systems using fault tree analysis for representing process uncertainty have also been proposed (Zhou et al., 2012; Wang et al., 2012).

Although expert systems have successfully been applied for fault detection and diagnosis, their limitations often overshadow their advantages. Some key limitation of an expert system approach include the system-specific nature of knowledge-based systems developed from expert rules, the difficulty in maintaining an expert system and the limited representation power of expert systems (Rich and Venkatasubramanian, 1987). In plant-wide implementations, expert systems would typically require a very high number of rules that could be distributed all over the plant information system, often with little or no logical linkage between them. Changes to such a distributed rule base could generate inconsistencies, resulting from the semantic impact of the changes on the rest of the system remaining hidden.

As with expert systems for fault detection and diagnosis, trend analysis (Cheung and Stephanopoulos, 1990; Janusz and Venkatasubramanian, 1991; Rengaswamy and Venkatasubramanian, 1995) can be used to explain process events, diagnose process malfunctions and predict future process states. For trend analysis, primitives are defined as the simple shapes into which the trend of a process variable can be demarcated. Ordered sets of these primitives, representing the trend of each process variable, are then monitored. Different variables are expected to each have a particular trend signature under normal operating process conditions. Under abnormal process conditions, it is expected that the different process variables will each exhibit trends that are characteristic of the particular abnormal condition, allowing different abnormal conditions to be mapped to characteristic trend signatures. For successful fault detection and diagnosis, any deviation in trend shape, duration or magnitude will be indicative of an abnormal process condition.

Sundarraman and Srinivasan (2003) proposed the use of a trend analysis-based approach for detecting events during process transitions. Since process transition can be characterised by the trends of key variables, any abnormality during a transition will results in the process variables evolving in a manner different from their normal trends. A comparison of the real-time trends of the key process variables with their normal trends, learned from historical process data, allow process transitions to be monitored. For real-time application to very large-scale plants, where the computational complexity of trend analysis increases with the number of sensors used for diagnosis, Maurya et al. (2005) proposed the use of PCA for dimensionality reduction prior to trend analysis. Whereas PCA is used to exploit the measurement redundancy, trend analysis is performed on the principal components for fault detection and diagnosis.

Trend analysis for fault detection and diagnosis has also been combined with interval-halving for non-linear trend extraction and fuzzy-matching for trend matching (Maurya et al., 2007b), signed directed graphs for improved diagnostic resolution (Dong et al., 2010) and using an adaptive window size for on-

Stellenbosch University http://scholar.sun.ac.za

line trend extraction (Maurya et al., 2010). More recently, Villez at al. (2013a) proposed the repeated application of Witkin's stability criterion during the feature selection step of trend analysis. This not only resulted in improved identification of the qualitative descriptions but also with regards to the locations in time of extrema and inflection points. Villez et al. (2013b) also presented a generalised method, based on a branch-and-bound algorithm for shape constraint spline fitting, for the analysis of a time series trends. Although the achieved identification is globally optimal, the proposed approach is regarded as too slow for real-time applications.

2.3 Data-based fault propagation paths

Data-based fault propagation paths deal with the capturing of process topology and causality from process data. With the aim of determining the root cause of a disturbance propagating widely through a large-scale plant, Bauer et al. (2007) turned to the application of transfer entropy, a quantitative process data-based method. Transfer entropy, through the exploitation of conditional probabilities (Schreiber, 2000), allows for the measurement of causality, and thus the establishment of cause and effect relationships, between different time series. In so doing, causal maps can be derived from process data collected during routine process operation, giving an indication as to the direction of propagation between different process measurements throughout the process. The application of direct transfer entropy was subsequently proposed (Duan et al., 2012) for the detection of direct causality between variables. Hou et al. (2010) proposed the inclusion of fault severity evaluation based on the key concepts of six-sigma in statistical quality control. Inclusion of a method for fault severity evaluation allows for different degrees of deviation to be estimated, providing useful information on the severity of a detected fault condition.

Similar to the use of transfer entropy, Bauer and Thornhill (2008) made use of the estimation of time delays between process variables to establish the cause and effect relationships between multiple variables. They proposed the use of the linear cross-correlation function for time delay estimation, quantifying the similarity of different process measurement over time, and a directionality index for determining the direction of the propagation, measuring the difference between the minimum and maximum values of the cross-correlation function. Likewise, Stockmann et al. (2012) proposed the use of the k nearest neighbor algorithm for the time delay estimation of non-linear systems, whereas Lu et al. (2012) proposed an approach based on the use time-delayed mutual information, not being tied to linear/non-linear properties of the system, for estimating both the causality direction and time-delay.

2.4 Signed digraphs & hierarchical decomposition

Although quantitative model-based approaches have been shown to be costly and complex to developed, there is a need for a reasoning tool which can be used to qualitatively model a system (Venkatasubramanian et al., 2003b). To this end, qualitative model-based approaches such as signed digraphs and hierarchical system decomposition not only allow the causal structure of a system to be captured but also to reason from it. Simply stated, a digraph is a graph, representing the cause-effect relations of a process, consisting of directed arcs leading from "cause" to "effect" nodes. Whereas events and variables are represented by nodes, edges represent the relationship between the nodes. By adding

Stellenbosch University http://scholar.sun.ac.za

a positive or negative sign to each of the directed arcs, signed digraphs are formed, providing a very efficient way of representing qualitative models graphically.

Iri et al. (1979) used a signed digraph for representing the influences among different elements in a chemical process with the aim of diagnosing system failures. The signed digraph is used to derive a cause-effect graph consisting of valid nodes, representing abnormal events or variables, and consistent arcs, potentially explaining the propagation of the fault through the process. The root cause of the fault is defined as the maximal strongly-connected components in the cause-effect graph and represents the pattern of the abnormality. Kramer and Palowitch (1987) derived logical rules from signed digraphs for fault diagnosis, explicitly representing the diagnostic criteria of the approach. Deriving a rule-base from signed digraphs not only allowed the solution to be tailored to reflect the best available knowledge of plant behaviour but also allowed integration with existing plant operating rules. Following this, Chang and Yu (1990) proposed various techniques for simplifying signed digraphs prior to deriving a rule-base for fault diagnosis. Various alternatives for improving the fault resolution of signed digraphs have also been suggested, some of which include the use of fuzzy logic (Han et al., 1994) or operational information in the form of a knowledge-base (Vendam and Venkatasubramanian, 1997). Vendam and Venkatasubramanian (1997) also demonstrated the use of signed digraphs for fault diagnosis from multiple origins.

For large-scale systems, Maurya et al. (2004) proposed combining unit specific signed digraphs with flow sheet connectivity. Unfortunately this led to the introduction of spurious solutions at flow sheet level which was subsequently reduced through the use of redundant equations and some quantitative process information. Signed digraphs have also been combined with qualitative trend analysis (Maurya et al., 2007a; Dong et al., 2010) and dynamic partial least squares (Ahn et al., 2008) for fault detection and diagnosis. When combined with qualitative trend analysis, signed digraphs are used to establish a possible candidate set of faults from which the root cause is determined through the application of trend analysis. This approach combines the completeness strength of signed digraphs with that of the diagnostic resolution strength of trend analysis. Similarly, when combined with dynamic PLS, signed digraphs are used to decompose the system into local diagnostic models with dynamic PLS being used to construct local models for each measured variable and in so doing improving the diagnostic resolution of the approach.

Recently, Wan et al. (2013) proposed the integration of signed digraph models with multivariate statistical process monitoring for root cause analysis of novel faults. As with certain prior hybrid solutions, signed digraphs are developed through the derivation of cause-effect relationships from process flow diagrams and subsequently used to identify candidate root nodes for root cause analysis. For each of the identified candidate root nodes, local primary residuals are generated and diagnosed using multivariate statistical process monitoring techniques. Lastly, local reconstructed residuals are generated from the fault direction matrices of the candidate root nodes, identifying the true root cause node. A common weakness identified in the application of signed digraphs for fault diagnosis is its weak diagnostic resolution.

Stellenbosch University http://scholar.sun.ac.za

Combining signed digraphs in hybrid solutions with other fault detection and diagnosis approaches does, however, seem to negate this limitation.

Hierarchical system decomposition deals with decomposing process systems into their respective process units, representing the functionality and structure of the process systems through the input-output relationships of the units. This is based on the observation where experienced plant operators break a problem into progressively smaller sub problems, seeking and prioritising several goals within each sub problem (Douglas, 1985). Whereas models at low levels of abstraction relate to a specific physical world, serving several purposes, models at higher levels of abstraction relate to a specific purpose that can be met by several physical groupings (Rasmussen, 1985). When traversing up an abstraction hierarchy the purpose of the respective process units with particular functions and how they serve higher level purposes are described. When traversing down an abstraction hierarchy the implementation of higher level purposes through process units with particular functions are described. For root cause analysis, Finch and Kramer (1987) decomposed a process system into a set of interacting subsystems and units where the failure of the purpose of a higher-level subsystem can be interpreted to be due to the failure of the function of one or more of the lower-level units. This not only allows the source subsystem/unit of a process fault to be quickly identified but also avoids any unnecessary detailed analysis during the initial stages of the diagnosis.

Lind (1991) proposed the hierarchical decomposition of a process system into a multilevel flow model, a normative description of the process system representing what it has been designed to do (goals), how it should do it (functions) and with what it should do it (physical components). Using this approach, a functional representation of the process in terms of how the physical components support sub-goals, how sub-goals support functions, and how the network of functions ensure the process goals are achieved can be obtained (Larsson, 1994). Similarly, Chen and Modarres (1992) used a goal tree-success tree model to hierarchically decompose a process system for fault diagnosis. The goal tree-success tree model shows a process system as a distributed network of unit operations and basic processes with nodes representing goals or functions. Whereas the higher levels of the model represents supervisory and management functions, at the lowest level plant hardware, human activities and process conditions are addressed. Modarres and Cheon (1999) subsequently proposed a function-centered approach, combining the goal tree-success tree model with a master logic diagram. In essence this combines the concept of structural and functional hierarchies through the use of “part-of” and “purpose-of” relationships into a “kind-of” hierarchy, based on the concept of classes, where an instance of a class will inherit all the properties and structural specifications of the class, but may still contain additional properties specific to that instance. Prasad et al. (1998) also used a hierarchical approach to organise diagnostic knowledge for a processing plant by primary processing systems, subsystems, components, behaviours and malfunction modes. In their approach different modules may use different methods to address specific local sub problems with development requiring readily available sources of knowledge, consideration given to common design and operating objectives of process plants and the inclusion of operating expertise and generic process characteristics.

Stellenbosch University http://scholar.sun.ac.za

Hierarchical decomposition in the form of multiblock methods has also received some attention, specifically with the aim of decomposing large processing systems, consisting of many measured variables, into conceptually meaningful blocks for the application of statistical data-based techniques (Slama, 1991; MacGregor et al., 1994; Wold et al., 1996; Westerhuis and Coenegracht, 1997). In the majority of approaches, the process is manually divided into blocks according to process domain knowledge, physical system constraints and/or process topology (Westerhuis et al., 1998; Qin et al., 2001; Bie and Wang, 2009; Zhang et al., 2010). Typically, logical subsystems are identified such that different processing steps or units are treated as independent blocks. More recently, Grbovic et al. (2012) proposed the use of sparse principal component analysis to partition the process by preserving the strongest correlations among sensors, thus producing a decentralised model with increased flexibility and fault tolerance compared to the centralised model alternative.

For large processes consisting of many processing units, each with large numbers of measured variables associated with them, fault detection and diagnosis systems can be quite complex and convoluted. The use of hierarchical system decomposition has been shown to provide effective modularity for structuring such large-scale diagnostic knowledge bases, allowing specific local problems to be addressed through the integration of different techniques (Prasad et al., 1998).

2.5 Process causality maps

It is evident that process performance monitoring has been extensively researched to date. However, no single method has been identified that exhibits all the desirable features required of a fault detection and diagnostic system as stipulated by Venkatasubramanian et al. (2003a). Additionally, Kiran et al. (2012) have identified various practical challenges regarding the development and implementation of accurate, customised fault monitoring and diagnostic models which include:

- Process variable selection for model construction.
- Fault detection and diagnosis technique reliability for highly non-linear processes.
- User friendliness of features facilitating the operators.

Several studies have reported the potential of, and benefits derived from, the use of hybrid systems for fault detection and diagnosis, combining fundamental process knowledge with data-base models (Van Lith et al., 2003; Venkatasubramanian et al., 2003c; Nandi et al., 2004; Van Sprang et al., 2005; Thornhill and Horch, 2007; Yang and Xiao, 2012). These hybrid systems employ different methods, complementing each other, to produce monitoring approaches with improved fault detection and diagnostic characteristics. This becomes especially important in large-scale industrial situations, where the integration of complementary features in order to overcome the limitations of individual solution strategies is much more crucial.

From a model-based perspective, it has been shown that the use of hierarchical system decomposition, as a qualitative model-based reasoning tool, can provide effective modularity for structuring large-scale diagnostic knowledge bases (Prasad et al., 1998). From a data-based perspective, statistical data-based approaches have been found to require very little modelling effort, making them relatively easy to

implement and more popular for large-scale industry applications (Chiang et al., 2001; Russell et al., 2000). The focus of this study is therefore on the development of a hybrid process performance monitoring system using process data-based approaches, derived from historical process data, supported by limited fundamental model-based approaches. Although the resulting hybrid approach may not meet all of the desirable characteristics of a diagnostic system as defined by Venkatasubramanian et al. (2003c), it should address some of the practical challenges highlighted by Kiran et al. (2012) and meet the current, most pressing needs of process performance monitoring for mineral processing plants.

Based on the ideas of some of the fundamental model-based approaches, especially hierarchical system decomposition, the development of process causality maps is proposed. Process causality maps are derived to describe the process being monitored both fundamentally and, through the integration of data-based techniques, statistically. Not only do process causality maps contextualise large amounts of data, but, since describing the interaction of the process units within the process, they also allow for the causal relationships between process units to be analysed. To this end an appreciation of causality and its potential application to process systems is needed.

2.5.1 Causality

Understanding the causal structure of a process is a fundamental capacity that allows process operators to control and predict the physical aspects of a process. At the most basic level process operators learn the causal relationships between process units through Pavlovian cue → effect experience. During such experience, cues (potential causes) precede effects and process operators having observed various cues and effects (or their absences), create their causal judgements of the causal strengths or efficacies of the various cues. Whereas process operators benefit from causal predictions about the process and subsequent corrective action, independent of any conscious thoughts they might have about causality, the success of their actions shows that some processes are sufficiently regular that even a rough rule of thumb can be a useful guide for controlling a process. The idea of causality is therefore very simplistic. Causality flows just one way without the back-and-forth possibilities of mutual influence.

Although the idea of causality is very simplistic, identifying causal relationship and understanding how cause and effect relations are inferred can be much more complex that the Pavlovian cue → effect experience allows for (Selltiz et al., 1959; Pearl, 2003; Hlaváčková-Schindler et al., 2007). Alternatives include chains of causes, multiple causes influencing one another; interactive causes, unobserved factors influencing both effects and observed potential causes, absence of prior knowledge or time order separating cause and effect, deterministic and probabilistic dependences, interventions that vary some factors while holding others constant, and uses of evidence involving both passive observation and interventions (Glymour, 2003). Also, especially when considering process systems, it is important to not only account for the forward material paths but also the recycle streams and information flow paths due to feedback control or even operator actions. Given this, it is obvious that establishing cause and effect relationships are not always straightforward. Causal claims should always be tested by competing alternatives or reversals of the claims. Often, perceived causal relations rely too much on assumption and too little on evidence.

Stellenbosch University http://scholar.sun.ac.za

Max Born (1949) stated three assumptions regarding cause and effect:

1. "*Causality* postulates that there are laws by which the occurrence of an entity Y of a certain class depends on the occurrence of an entity X of another class, where the word *entity* means any physical object, phenomenon, situation, or event. X is called the cause, Y the effect."
2. "*Antecedence* postulates that the cause must be prior to, or at least simultaneous with, the effect."
3. "*Contiguity* postulates that cause and effect must be in spatial contact or connected by a chain of intermediate things in contact."

He concluded that the "overwhelming evidence of causality with all its attributes in the realm of ordinary experience is satisfactorily explained by the statistical laws of large numbers". These statistical laws are the formula to which real events truly conform and can be observed, tested and verified.

All observable processes are law governed to some extent, although the laws may not always be known in complete detail (Sowa, 1999). Law governed processes assume that some things are predictable, while others aren't. Such processes form a continuous range of possibilities between totally random processes, where nothing is predictable, and totally deterministic processes, where everything is completely predictable. Laws, therefore, allow process operators to make predictions about the future operation of processes. Relating predictability to cause and effect, Wiener (1956) defined a mathematical definition for causality: X could be defined to cause Y if the predictability of Y is improved by incorporating information about X. Granger (1969) extended this definition into the experimental practice, formalising the prediction idea in the context of linear regression models: X is said to have a causal influence on Y if the variance of the autoregressive prediction error of Y at the present time is reduced by inclusion of past measurements of X.

Constraints also play a vital role in establishing causal relationships within a process. A constraint on a process is a relation among the functions of that process that limits their range of permissible values. A causal constraint limits the causal influences to a region called the cone of causal influence (named after the light cone as introduced by Albert Einstein in the theory of relativity), since anything outside the cone cannot influence or be influenced by anything that happens in the present (Sowa, 1999). Laws place constraints on the variability of the process functions over time, limiting the uncertainty in the predictability of a process. For law governed processes the question is not whether they are predictable, but whether the constraints can be tightened to limit the range of uncertainty.

Although physical processes are continuous, for most methods of reasoning, measurement and computation, discrete approximations of the process is required. Discrete processes can be seen as a combination of states (conditions persisting for some duration) and events (changes occurring between states). For every event type there exist a fixed number of input state types (preconditions) and a fixed number of output state types (post conditions). The combination of preconditions and post conditions forms the signature of the event type.

Stellenbosch University http://scholar.sun.ac.za

Predictions are performed through the inverse mapping from the discrete approximation to the continuous process. However, due to the limited accuracy of all measuring instruments and the limited storage capacity of even the largest digital computers, predicted values cannot correspond exactly to the continuous physical values, they can only approximate them. Yet, if the difference between the actual value and the predicted value is sufficiently small, a predicted value may be adequate for a particular application. For an engineer, the goal is to find an adequate approximation to solve the problem at hand. However, it is important to remember that even the best models are approximations to a limited aspect of the world for a specific purpose (Sowa, 1999).

2.5.2 Hierarchical system decomposition

Hierarchical system decomposition provides the framework for deriving process causality maps from a process system. It not only allows for the retention of process flow topology and connectivity information via structural decomposition, but also via functional decomposition for the use of causal and fundamental process and unit specific information for variable selection and results interpretation.

Hierarchical system decomposition deals with decomposing process systems into their respective process units, representing the functionality and structure of the process systems through the input-output relationships of the units. This is based on the observation where experienced plant operators break a problem into progressively smaller sub problems, seeking and prioritising several goals within each sub problem (Douglas, 1985). It has been found that the structuring of a process system into different level typically can typically be done along two dimensions (Rasmussen, 1985): structural and functional. Structural decomposition is where the process system is seen to consist of various related components at different levels of physical aggregation. Functional decomposition is where the physical implementation of the process functions is maintained in the hierarchical representation. Although not always the case, very often a change in the level of structural decomposition is coupled with a change in the level of functional decomposition.

For the functional decomposition of a process system, the process- and equipment independent functional properties of the system are represented in several levels of functional abstraction with the number of levels required between the material physical implementation and the ultimate purpose of the system being dependent on the reason of the decomposition as well as the type of process system. At the lowest level of decomposition, the physical form of the system as defined by the material configuration is represented. Moving higher up the abstraction hierarchy, the physical processes or functions of the various components and systems are represented, until at the highest levels, only general concepts are used with which to represent the functional properties of the process system with no reference made to the physical process or equipment (Rasmussen, 1985). When moving up the abstraction hierarchy, information regarding the physical and material properties of the system is not simply removed, but information is also added regarding the principles governing the cofunction of the lower level functions and elements.

Stellenbosch University http://scholar.sun.ac.za

Whereas models at low levels of abstraction relate to a specific physical world, serving several purposes, models at higher levels of abstraction relate to a specific purpose that can be met by several physical groupings. From this it is evident that bottom-up an abstraction hierarchy describes the purpose of the respective process units with particular functions and how they serve higher level purposes while top-down an abstraction hierarchy describes how higher level purposes are implemented through process units with particular functions. Hierarchical decomposition of a process system therefore not only transforms a specific problem to a level at which the solution is more readily available, but also serves as a powerful functional reasoning tool, allowing interaction at any level of the abstraction hierarchy (Rasmussen, 1985).

For fault administration (including both recognition and correction), Chen and Modarres (1992) used hierarchical system decomposition in the form of a goal tree-success tree (GTST) to represent deep-knowledge of the process. A major objective of this approach is the formulation of a general knowledge representation depicting a complex physical system in a simple manner. Included in this, the representation should provide sufficient data structures to precisely describe the object while rigorously supporting the logical relationship between each object and its function(s). In summary, this approach not only provides the rigorous definition of logical connections in a system structure and function, but also the process topology.

The GTST model shows a processing plant as a distributed network of unit operations and basic processes with embedded monitoring, instrumentation and control systems as well as higher level management and supervisory functions. The nodes in the model represent goals or functions (in the case of physical systems); the top plant goal/objective being explicitly defined first, with progressively more detailed lower levels as the goal tree is decomposed vertically downward from this goal/objective. For any goal it must be possible to define explicitly how the specific goal/sub-goal is satisfied while simultaneously being able to define explicitly why the specific goal/sub-goal must be satisfied. At the lowest level of the goal tree plant hardware, human activities and process conditions are addressed, representing the possible combinations of operating components which achieve a goal in the goal tree (Figure 3). For this a thorough understanding of the physical aspects of the process and its properties is required.

Stellenbosch University http://scholar.sun.ac.za

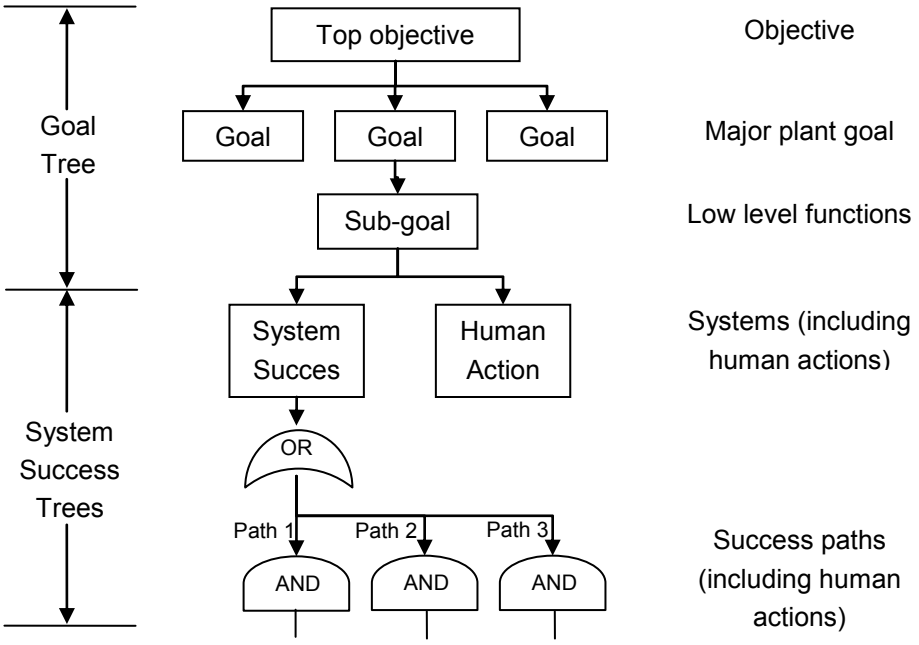

Figure 3: **Typical structure (frame-based representation) depicting the hierarchical relationship of a GTST model (adapted from Chen and Modarres, 1992)**

The top goal is hypothesized to have failed whenever one of the above SPC criteria is met. All sub-goals of the top goal should subsequently be investigated to determine the source of the problem. It should be noted that the correct identification of the problem could be hampered by uncertain evidences or conclusions arising from uncertainty in data, data interpretation and rules of inference. In reality, some information is always unavailable to support the desired logical inferences of fault recognition. To account for this, Chen and Modarres (1992) used the Bayesian theorem to handle evidence and deduction uncertainties for faulty recognition. Furthermore, sensor failure may lead to misdiagnosis requiring comprehensive data validation of all sensor data before a diagnostic conclusion is reached.

Lind (1991) proposed the hierarchical decomposition of a process system into a multilevel flow model, a normative description of the process system representing what it has been designed to do (goals), how it should do it (functions) and with what it should do it (physical components). Using this approach, a functional representation of the process in terms of how the physical components support sub-goals, how sub-goals support functions, and how the network of functions ensure the process goals are achieved can be obtained (Larsson, 1994). Within the functional representation, different types of relations between the goals, functions and physical components were defined. When connecting a set of functions to a goal with the aim of implying that the functions are used to obtain the goal, an achieve relation is used. When connecting a goal to a function with the aim of implying that the function only becomes available should the goal be fulfilled, a condition relation is used. When connecting a physical component to a function with aim of implying that the function is implemented or realised through the use of the physical

Stellenbosch University http://scholar.sun.ac.za

component, a realise relation is used. Whereas goals are defined as either production, safety or economy related, flow structures are treated as either mass, energy or information related.

For fault diagnosis, each of the flow functions within the multilevel flow model could be in a state of normal, working of faulty. In order to establish the relationship between the failure states of the physical components and that of the different flow functions, the failure state of the function need to be investigated through a test to be performed, a check to be done, or a question to be asked. A search of the model graph is ultimately used for fault diagnosis (Larsson, 1994):

1. A goal/sub-goal for diagnosis is selected.
2. Achieve relations are used to search from the goal/sub-goal into the connected network of flow functions.
3. Realise relations are used to determine if the physical components corresponding to each of the flow functions are currently realising the functions.
4. For a faulty flow function conditioned by a sub-goal, or if the flow function cannot be checked, the search is repeated from step 2.
5. For a working flow function the searched is deemed completed and that part of the sub-tree skipped.

Larsson (1994) found that the process of implementing and maintaining knowledge databases for fault diagnosis using multilevel flow models were much more efficient than rule-based expert systems due to the graphical nature of multilevel flow models. The graphical nature further made inconsistencies in the database impossible, thus guaranteeing consistency within the model.

Prasad et al. (1998) also used a hierarchical approach to organise diagnostic knowledge for a processing plant by primary processing systems, subsystems, components, behaviours and malfunction modes. In their approach different modules may use different methods to address specific local sub problems with development requiring readily available sources of knowledge, consideration given to common design and operating objectives of process plants and the inclusion of operating expertise and generic process characteristics.

A hierarchical approach provides a platform for the decomposition of the process with which to manage and distribute diagnostic decision making into a set of focused but coordinated problem solving modules with the ability to associate different diagnostic methods to different problem solving modules. Furthermore, engineered processes typically display hierarchical whole-part organisation, allowing the efficient modularisation of the diagnostic knowledge-base while simultaneously mirroring the structure of modules of the process system to be diagnosed. Each node in the hierarchy is used to define a localised diagnostic method specialised for establishing whether or not a relevant malfunction exists at that node, with the tip nodes of the hierarchy representing malfunction modes of specific equipment or instrumentation. Evaluation of a malfunction follows a top down approach down the hierarchical model of the process from a general to a specific malfunction.

Stellenbosch University http://scholar.sun.ac.za

Prasad et al. (1998) defined four types of available knowledge that should be captured:

- Structural – physical equipment and connections, useful for structuring process hierarchy
- Functional – intended functions and operating conditions, useful for structuring process hierarchy
- Malfunction – what and how things can go wrong
- Behavioural – causal consequences of conditions, crossing functional and structural boundaries

Causal knowledge is especially required to isolate and identify faults in weakly instrumented plants where malfunction symptoms may only appear causally, well downstream from the initial fault.

For monitoring purposes, three distinct plant operating objects (production, product quality and safety) have been identified (Prasad et al., 1998). For the production objective, knowledge is structured into primary process systems, subsystems, equipment items and modes of failure. For primary process systems boundaries are drawn around related equipment groups whereas subsystems are modules that have been engineered into the process design so they present meaningful groupings from a process operation or human-interaction perspective. For large processes consisting of many processing units, each with large numbers of measured variables associated with them, fault detection and diagnosis systems can be quite complex and convoluted. The use of hierarchical system decomposition has been shown to provide effective modularity for structuring such large-scale diagnostic knowledge bases, allowing specific local problems to be addressed through the integration of different techniques (Prasad et al., 1998).

Modarres and Cheon (1999) subsequently proposed a function-centered approach, combining the goal tree-success tree model with a master logic diagram. Given the structural decomposition of a process system, the relationships between various objects are described as a “part of” hierarchy: each object in the hierarchy being a “part of” another object situated at a higher level of abstraction. Given the functional decomposition of a process system, the relationships between various objects are described as a “purpose of” hierarchy: association with objects in the hierarchy being an arbitrary process that is highly influenced by the “purpose of” the decomposition. When representing the classes of objects in a hierarchy, indicating the hierarchy of the classes of objects, the relationship between the various classes are described as a “kind-of” hierarchy. In essence the function-centered approach combines the concept of structural and functional hierarchies through the use of “part-of” and “purpose-of” relationships into a “kind-of” hierarchy, based on the concept of classes, where an instance of a class, an object, will inherit all the properties and structural specifications of the class, but may still contain additional properties specific to that instance. The resulting functional-centered hierarchy is conceptually represented in the form of a goal tree-success tree model and combined with a master logic diagram. With the highest level of the master logic diagram forming the lowest level of the goal tree-success tree, the combined representation can be used to model complex plants.

Stellenbosch University http://scholar.sun.ac.za

2.5.3 Process causality maps

Based on this overview, general rules can be derived for the construction of process causality maps. Process causality maps can be seen as consisting of process flow diagrams with unit specific causality maps hanging off each process area/unit (Figure 4). At the highest level the process causality map represents the process as a single functional node with process performance measures for process (production and quality), equipment and/or control. The process is further decomposed into areas and units in subsequent layers of the process causality map, following the general layout of the process flow diagram, accounting not only for the forward material paths but also recycle streams. Relevant process performance measures, in the form of data-based techniques, are attached to the various functional nodes within the process causality map, allowing for top-down root cause analysis to be performed as well as interaction or intervention with the model at any level. At the lowest level of the process causality map individual instrument units are found with process performance measures for data validation and instrument failure detection.

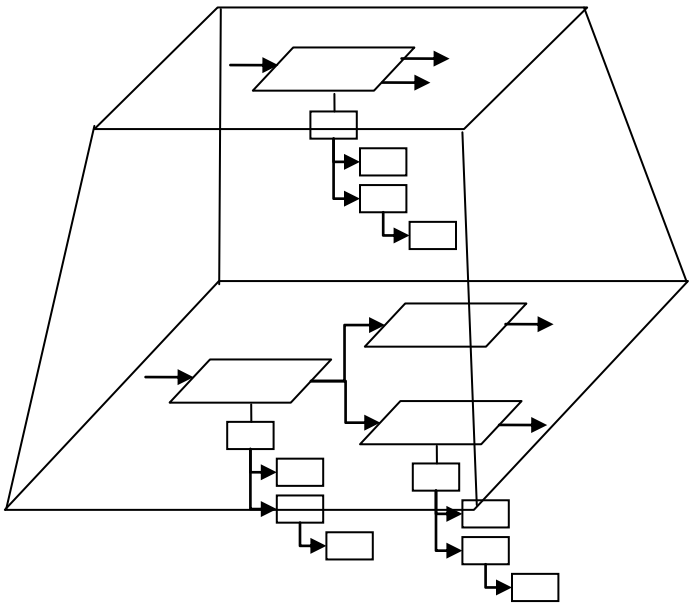

Figure 4: **Generic process causality map depicting a process flow diagram as multiple layers with unit specific causality maps hanging off each process area/unit**

The aim of structuring the process as such is not to over analyse the process but rather to treat the process as a series of interconnected major process units with each process unit being functionally described by the fundamental physics of that process unit. Division of the process into areas and units are based on their location within the process flow, the interactions and interconnections between the areas and units, the function and purpose of the major units and their mutual dependence during operation. All process data (including high frequency online process data and low frequency offline

Stellenbosch University http://scholar.sun.ac.za

laboratory data) are thus grouped logically into functional nodes and combined with the relevant metadata.

At the various functional nodes, process performance measures, in the form of data-based techniques, are attached which suit the fundamental drivers of the process units. Process performance measures can be anything from single values or simple time series plots to complex custom statistical data analysis techniques used to monitor the process performance with. These measures are not only used to describe the performance of the various process units but they are also used to monitor the process equipment and control solutions (which inherently affects process performance).

When combining relevant functional nodes hierarchically, a unit specific causality map is formed (Figure 5). The hierarchical nature of the unit specific causality map is a combination of how various process units and process measurements relate to one another. For any two functional nodes to be connected to one another within a unit specific causality map, a mathematical or relational relationship needs to exist between the two functional nodes. The relationship could be in the form of data being condensed from the bottom functional node to the upper functional node through the calculation of a simple process performance measure or the association of information due to the fundamental physical relationship between the functional nodes. If it is not possible to define a relationship between two functional nodes, it is impossible to determine the root cause at the lower functional node due to a shift in the process performance at a higher functional node in the unit specific causality map. Constructing such a unit specific causality map, therefore, not only requires a sound knowledge of how the various process variables relate to one another, but also a good fundamental understanding of how the various process units operate and interact with each another. Unit specific causality maps are a culmination of process, equipment and control knowledge.

Stellenbosch University http://scholar.sun.ac.za

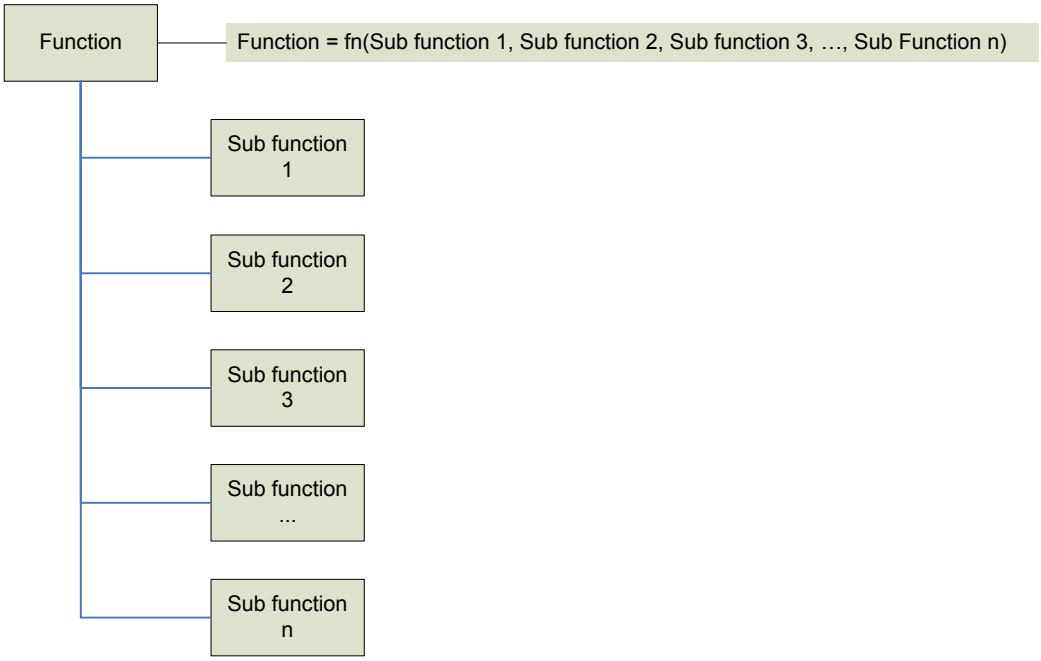

Figure 5: **Generic unit specific causality map where the top level of the tree (Function) is the primary function of the process unit and the sub functions represent the important identifiable parameters that contribute to the function**

The use of process causality maps, combined with data-based methods, therefore addresses the following desirable characteristics (highlighted below) a diagnostic system should possess (Venkatasubramanian et al., 2003a):

1. **Quick detection and diagnosis**
2. **Isolability**
3. **Robustness**
4. **Novelty identifiability**
5. Classification error estimate
6. **Adaptability**
7. **Explanation facility**
8. **Modelling requirements**
9. **Storage and computational requirements**
10. **Multiple fault identifiability**

Of these, it is expected that the proposed process causality maps will contribute mostly in terms of explanation facility.

Stellenbosch University http://scholar.sun.ac.za

3 Statistical data-based fault detection

Various statistical methods exist for the analysis, monitoring and diagnosis of process operating performance over time with those based within the SPC framework having received much attention. Numerous SPC techniques exist, applicable to a wide range of monitoring scenarios covering univariate, multivariate, static and dynamic monitoring of linear and non-linear processes.

Statistical data-based fault detection techniques all make use of a mathematical representation of the process based on historical process data. The use of data-based models is especially advantageous when encountering high process non-linearity, high dimensionality or high process complexity. These data-based methods use historical process data to verify whether or not a process is within a "state of statistical control" while still allowing "common-cause" variation (variation that affects the process all the time and is essentially unavoidable within the current process). Typically a statistical model is built using historical process data when the process was in a state of normal operating condition (NOC). Any periods containing variations arising from special events that one would like to detect in the future are omitted, but can be used to build individual event/fault models. New process data is then compared to this NOC model in order to detect a change in the system. When the new process data is compared to the event/fault models specific events can be identified.

Changes within the system can be detected using either univariate or multivariate data. Conventional univariate control charts, such as Shewhart, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charts usually include both warning limits (target value ± 2σ) and control limits (target value ± 3σ), where σ is the standard deviation of the time series data. Heuristic rules (Nijhuis et al., 1997) can then be used to determine if an out-of-control situation exists:

- Any observations outside of the control limits.
- Two or more consecutive observations outside the warning limits.
- Four or more consecutive observations outside the 1σ limits.
- Eight or more consecutive observations moving in either an upward or downward direction.
- Any unusual or non-random pattern in the data.

However, these charts commonly only permit investigation into the magnitude of deviation of any one variable, independently of all other variables, at a given time, often resulting in inaccurate, delayed conclusions being drawn (Kourti and MacGregor, 1995). Essential process information therefore may not necessarily lie within any individual process variable (univariate), but also in how the variables change with respect to one another (multivariate). Multivariate statistical projection and regression methods such as principal component analysis (PCA) or partial least squares (PLS) overcome many of these shortfalls.

This chapter deals with the theory behind many of the statistical process monitoring techniques available for statistical data-based fault detection, focussing on those within the SPC domain. The selected techniques are not an exhaustive list of available techniques, but have been chosen to be illustrative of what is available.

Stellenbosch University http://scholar.sun.ac.za

3.1 Univariate statistical process monitoring

For statistical process monitoring (SPM), conventional univariate control charts all monitor the deviation, or variation thereof, between a single process value and its associated target value. The most frequently used univariate control charts are the Shewhart, CUSUM and EWMA charts.

3.1.1 The Shewhart chart

The classical Shewhart chart (Shewhart, 1931) consists of a sequential time plot of process data, a target value and upper and lower control limits set at ± 3σ from the target value. In essence, it is a graphical method for testing whether or not the observed process measurement differs significantly from its desired target value. The Shewhart chart has been found to be effective in detecting large changes in the mean of the process variable being monitored, it being much less sensitive to shifts of small magnitude.

A multivariate Shewhart chart can be constructed by plotting PCA or PLS scores or the Hotelling's T^2 statistic based on these scores on a Shewhart chart. Evolution of the process over time with regard to the PCA or PLS model is given by a Shewhart chart of the PCA or PLS scores whereas an estimate of how far away an observation is from the centre of the PCA or PLS model is given by a Shewhart chart of Hotelling's T^2 statistic (showing a summary of all the scores).

3.1.2 The CUSUM chart

The CUSUM chart (Page, 1954) was developed for detecting small changes in the mean of a process variable. It is a plot of the cumulative sum of the deviation of the process data being monitored from its desired target value. The CUSUM chart can be seen as a variation on the Shewhart chart where the cumulative sum of the deviations is monitored instead of the deviations themselves. Another important difference to note is that whereas the Shewhart method focuses only on the most current information the CUSUM method pays equal attention to the entire data history.

A multivariate CUSUM chart can be constructed by plotting PCA or PLS scores on a CUSUM chart. Typically a cumulative sum adding the deviation on the high side scores larger than the target value and a cumulative sum adding the deviation on the low side scores smaller than the target value will be monitored.

3.1.3 The EWMA chart

The EWMA method (Roberts, 1959) can be regarded as a moving average technique, also making it a dynamic method, that falls somewhere between the Shewhart and CUSUM methods (Ogunnaike and Ray, 1994). It uses a first-order digital filter to take the history of the data into account (similar to the CUSUM method), typical for a process with memory and drift, and the filtered values are subsequently monitored (similar to the Shewhart method). The parameter for determining the memory length strongly influences the EWMA variance. Its choice is very important in determining the methods' sensitivity to detecting shifts in the mean of the process variable being monitored. The EWMA chart can also be

extended to the multivariate case (Lowry et al., 1992). A multivariate EWMA chart can be constructed by plotting PCA or PLS scores on a EWMA chart.

3.2 Basic multivariate statistical process monitoring

PCA and PLS treats all process variables simultaneously. In short they extract information on how all the process variables are behaving relative to one another and compress the information in a manner which retains essential information. PCA and PLS is therefore ideal for handling high dimensional and correlated variables, has a built-in noise-filter and facilitates the graphical interpretation of multivariate results.

3.2.1 Principal component analysis

Conventional PCA finds linear combinations of the variables in a data set X, consisting of n rows (observations) and p columns (variables), through eigenvector decomposition of the correlation matrix of the variables (Wise and Gallagher, 1996). Prior to PCA all the variables in X must first be auto-scaled. Auto-scaling is required so as to avoid having important variables whose magnitudes are small from being overshadowed by less important, but larger magnitude variables. During auto-scaling the data is adjusted to zero mean and unit variance. After auto-scaling, the covariance matrix of X can be defined as:

$$\mathrm{cov}(X) = \frac{X^T X}{n-1} \tag{3-1}$$

X is consequently decomposed as the sum of the outer product of vectors t_i and l_i plus a residual matrix E:

$$X = t_1 l_1^T + t_2 l_2^T + \ldots + t_a l_a^T + E \tag{3-2}$$

where a is less than or equal to the smallest dimension of X. The t_i vectors (scores) contain information on how the observations relate to one another, and the l_i vectors (loadings) contain information on how the variables relate to one another. The l_i vectors are eigenvectors of the covariance matrix:

$$\mathrm{cov}(X) l_i = \lambda_i l_i \tag{3-3}$$

where λ_i is the eigenvalue associated with the eigenvector l_i. For any t_i, l_i pair, λ_i is a measure of the amount of variance (information) described by the t_i, l_i pair. The t_i, l_i pairs are usually arranged in

Stellenbosch University http://scholar.sun.ac.za

descending order according to its associated λ_i. The first t_i, l_i pair then captures the largest amount of variance of any pair in the decomposition and each subsequent pair captures the largest possible amount of variance remaining at that step. It should also be noted that the score vector t_i is a linear combination of the data set X defined by l_i:

$$Xl_i = t_i \tag{3-4}$$

Following PCA, it is generally found that the data can be adequately described using far fewer factors than the original variables. The score variables produced are also usually more normally distributed than the original variables themselves, except where they are associated with a controlled property of the system. Various statistics exist with which to monitor/validate the PCA model. The most common of these is the Hotelling's T^2 statistic, giving an indication as to the variation within the model, and the lack of model fit statistic, Q.

3.2.2 Partial least squares

For a given input, X, and output, Y, data set, PLS regression attempts to find factors that not only capture the greatest amount of variance in the predictor variables but is also the best at correlating the predicted explanatory variables with predicted response variables. PLS can therefore be said to attempt to maximise covariance (Wise and Gallagher, 1996).

In PLS, similar to PCA, the X and Y data sets are auto-scaled and then decomposed:

$$X = t_1 l_1^T + t_2 l_2^T + \ldots + t_a l_a^T + E \tag{3-5}$$

$$Y = u_1 q_1^T + u_2 q_2^T + \ldots + u_a q_a^T + G \tag{3-6}$$

The latent vectors, t_a, are computed from the data for each PLS dimension such that the linear combination of the X and Y variables, respectively defined by the latent variables $t_a = l_a^T x$ and $u_a = q_a^T y$, maximises the covariance between X and Y that is explained at each dimension. The most instructive method of calculating the PLS model parameters is probably the non-iterative partial least squares (NIPALS) algorithm. The NIPALS algorithm calculates not only scores and loadings, but also an additional set of vectors know as weights, Z, for maintaining orthogonal scores. In addition to this, a vector of "inner-relationship" coefficients, b, relating the X- and Y-block scores, are also computed.

Stellenbosch University http://scholar.sun.ac.za

The PLS decomposition is started by selecting a column of Y, y_j, as the starting estimate for u_1. For the X data block:

1. $$z_1 = \frac{X^T u_1}{\left\| X^T u_1 \right\|} \tag{3-7}$$

2. $$t_1 = X z_1 \tag{3-8}$$

Then, for the Y data block:

3. $$q_1 = \frac{u_1^T t_1}{\left\| u_1^T t_1 \right\|} \tag{3-9}$$

4. $$u_1 = Y q_1 \tag{3-10}$$

Convergence is checked by comparing the calculated t_1 from step 2, with the one from the previous iteration. Repeat steps 1 to 4 until the calculated t_1 are equal within rounding error using u_1 as calculated in step 4. When they are equal within rounding error, calculate the X data block loadings and rescale the scores and weights accordingly:

5. $$l_1 = \frac{X^T t_1}{\left\| t_1^T t_1 \right\|} \tag{3-11}$$

6. $$l_{1new} = \frac{l_{1old}}{\left\| l_{1old} \right\|} \tag{3-12}$$

7. $$t_{1new} = t_{1old} \left\| l_{1old} \right\| \tag{3-13}$$

8. $$z_{1new} = z_{1old} \left\| l_{1old} \right\| \tag{3-14}$$

Next, the regression coefficient b is found for the inner relation:

Stellenbosch University http://scholar.sun.ac.za

9. $b_1 = \frac{u_1^T t_1}{t_1^T t_1}$ (3-15)

Once the scores and loadings have been determined for the first factor (called a latent variable in PLS), the X - and Y -block residuals are calculated:

10. $E_1 = X\text{-}t_1 l_1^T$ (3-16)

11. $G_1 = Y\text{-}b_1 u_1 q_1^T$ (3-17)

The entire procedure is now repeated from step 1 for the next latent variable. X and Y are replaced with their residuals E_1 and G_1 respectively, and all subscripts are incremented by 1.

The PLS calculated scores and loadings can be interpreted in terms of PCA scores and loadings that have been rotated to be more relevant for predicting y .

For a mineral processing plant Hodouin et al. (1993) applied both PCA and PLS to monitor 44 process variables and 23 calculated variables averaged hourly over 350 hours. The system was first decomposed into four sub-systems namely the grinding circuit (GC_{MPP}), the flotation circuit (FC_{MPP}), the rougher/scavenger flotation stage (RS_{MPP}) and the cleaning/regrinding flotation stage (CR_{MPP}). The variables were then grouped into matrices of observation which were used to construct input and output blocks for each of the sub-systems.

Systematic errors (accounting for bias in the data) in the subsequent data set were corrected for and followed by a mass balance to reconcile the data, correcting random errors (accounting for dispersion in the data). For each sub-system PCA was applied to the matrix formed by combining the input and output blocks of the relevant sub-system and the results analysed. PLS was also applied to the input and output blocks for each of the sub-systems using either a multi-input-single-output or multi-input-multi-output approach allowing various relationships in the data to be analysed.

3.2.3 T^2-chart

The T^2-chart is a multivariate chart based on Hotelling's T^2 statistic (Hotelling, 1947). This chart is plotted based on the first a principal components (PCs) (Kourti and MacGregor, 1995), where:

$$T_a^2 = \sum_{i=1}^{a} \frac{t_i^2}{\lambda_{t_i}} \quad (3\text{-}18)$$

Stellenbosch University http://scholar.sun.ac.za

and λ_{t_i} is the estimated variance of t_i according to the PCA model based on historical data. Each t_i^2 is scaled by the reciprocal of its associated variance, λ_{t_i}, in order to ensure that each PC term plays an equal role in the computation of T^2, irrespective of the amount of variance it explains in the data set X.

Using historical data, the upper confidence limit (UCL) for the T^2-chart is given by (Kourti and MacGregor, 1995):

$$T_{UCL}^2 = \frac{(n-1)(n+1)a}{a(n-a)} F_\alpha(a, n-a) \qquad (3\text{-}19)$$

where n is the number of observations in the data set, a is the number of PCs retained in the PCA model and $F_\alpha(a, n-a)$ is the upper $100\alpha\%$ critical point of the F-distribution with a and $n-a$ degrees of freedom. The 95% UCL is calculated using $\alpha = 0.05$ and the 99% UCL is calculated using $\alpha = 0.01$.

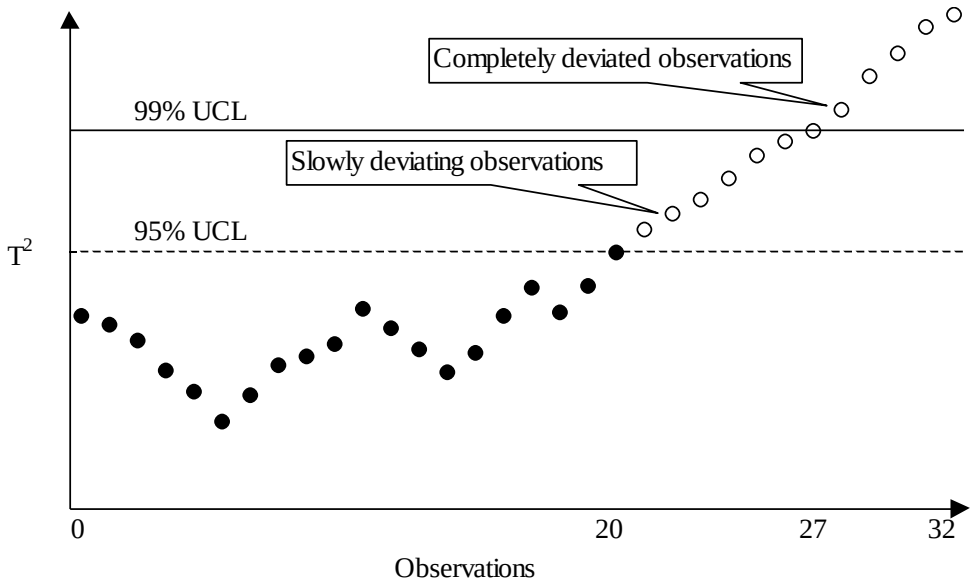

Figure 6: ***T*²-chart indicating abnormal process behaviour resulting from when the variation in the variables is greater than can be explained by "common-cause" (adapted from Kharva, 2001)**

The T^2-chart (Figure 6) provides a test for deviations in the variables that are of greatest importance to the variance of the data set X. It will, however, only detect whether or not the variation in the variables in the plane of the first a PCs are greater than can be explained by "common-cause".

Stellenbosch University http://scholar.sun.ac.za

In order to identify the cause of the deviations in the variables, variable contributions can be calculated. For Hotelling's T^2 statistic, variable contributions are computed as $\left(t\sqrt{\Lambda^{-1}}L'\right)$, where Λ is a diagonal matrix with diagonal elements equal to eigenvalues and off-diagonal elements equal to zero. The score vector t can also be substituted by the difference of two x vectors multiplied by the loading matrix L, shedding light on the difference between two points in the principal component subspace (Teppola et al., 1998).

If the variables stay within the same min-max limits as the reference data set used to develop the PCA model, and the (internal) relationships between the variables change, new PCs will appear and the new observations will move away from the hyper-plane defined by the reference PCA model. This change can be detected using a SPE-chart.

3.2.4 SPE-chart

The Squared Prediction Error (SPE) chart is a multivariate chart based on the Q statistic. The SPE-chart measures the amount of variation in each observation not captured by the a principal components retained in the PCA model. This is accomplished by calculating the SPE of the residuals of new observations (Kourti and MacGregor, 1995):

$$SPE_x = \sum_{i=1}^{q}\left(x_{new,i} - \hat{x}_{new,i}\right)^2 \qquad (3\text{-}20)$$

where $x_{new,i}$ is a PCA input, and $\hat{x}_{new,i}$ is the prediction of $x_{new,i}$ from the PCA model. Using historical data, the UCL for the SPE-chart is given by (Nomikos and MacGregor, 1995):

$$SPE_{UCL} = g\chi^2_{h,\alpha} \qquad (3\text{-}21)$$

where $g = \upsilon/2\mu$, the weight of the weighted chi-squared distribution, and $h = 2\mu^2/\upsilon$, the degrees of freedom of the weighted chi-squared distribution, with μ the mean and υ the variance of the SPE historical data set, at significance level α. SPE confidence limits can be established for the residuals of the data set X or for the residuals of the individual variables. Analogous to the T^2-control chart, the 95% UCL is calculated using $\alpha = 0.05$ and the 99% UCL is calculated using $\alpha = 0.01$.

Stellenbosch University http://scholar.sun.ac.za

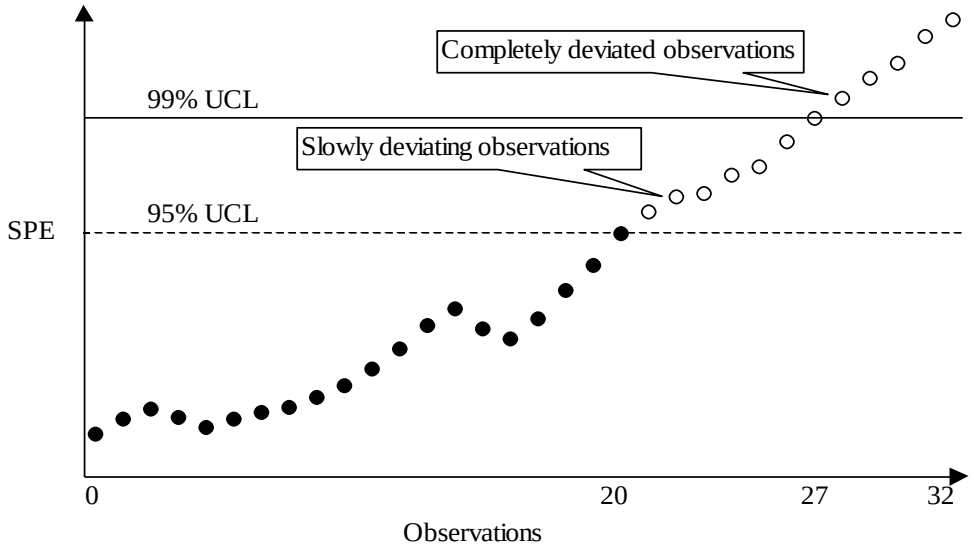

Figure 7: SPE **-chart indicating abnormal process behaviour resulting from when the relationship among the variables being monitored or their dynamic patterns change (adapted from Kharva, 2001)**

The SPE -chart (Figure 7) indicates abnormalities resulting from when the relationship among variables or their dynamic patterns change, indicating a breakdown of the correlation among the variables due to the potential existence of an event. Under these conditions, the PCA model can no longer explain the new correlation among the measurements and the SPE will increase.

For the SPE, variable contributions are computed as a vector of squared prediction errors $\left(x_{new,i} - \hat{x}_{new,i}\right)^2$ in which each of the vector elements is the corresponding variable's contribution to the prediction error (Westerhuis et al., 2000).

3.2.5 Score-chart

The score-control chart is a multivariate chart constructed by plotting the scores, t_i, obtained from applying PCA to the data set X in either two or three dimensions. If more than three PCs were retained in the PCA model, plotting various combinations of t_i could result in valuable process information being extracted.

Similar to the other control charts, confidence limits can be placed on the process scores. Wise and Gallagher (1996) suggested that the confidence limits on the process scores be established based either on judgement concerning the desired process operating limits or using more sophisticated time series

Stellenbosch University http://scholar.sun.ac.za

modelling techniques. Should the data be normally distributed, the T-distribution could be used to calculate the UCL for the score-chart.

Using historical data, the UCL for the score-chart is given by:

$$T_{UCL} = \sigma_i \hat{t}_{\alpha/2,n-1} \tag{3-22}$$

where σ_i is the standard deviation of the score t_i and $\hat{t}_{\alpha/2}$ is the upper $\alpha/2$ point of the T-distribution with $n-1$ degrees of freedom. The 95% UCL is calculated using $\alpha = 0.05$ and the 99% UCL is calculated using $\alpha = 0.01$.

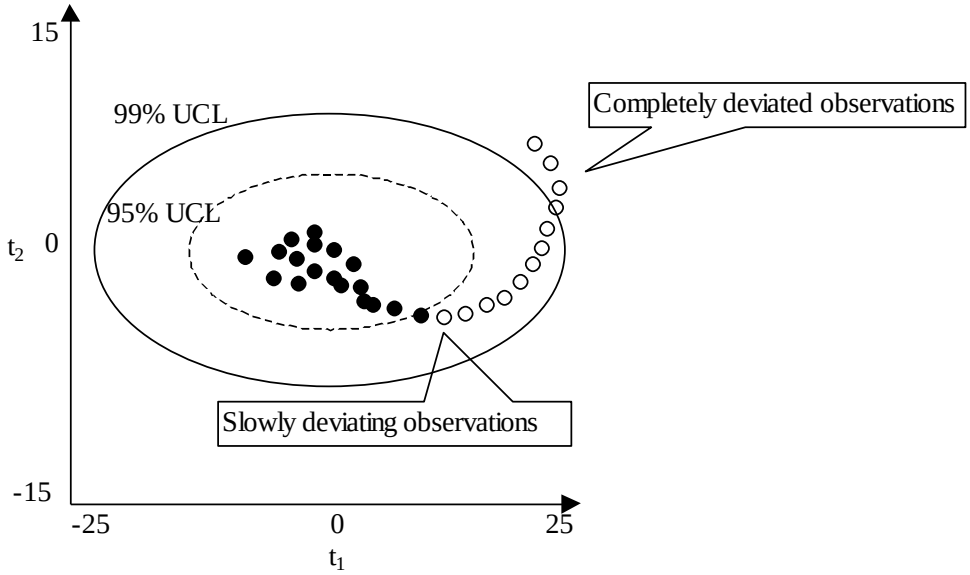

Figure 8: **Score-control chart indicating abnormal process behaviour**

The score-chart (Figure 8) indicates abnormalities resulting from several variables having a larger than normal change, whereas the basic relationship among the variables and their dynamic patterns do not change. Under these conditions the principal scores will move outside the region over which the PCA model was developed while the SPE will remain small. Combination score-SPE-charts are also possible.

Stellenbosch University http://scholar.sun.ac.za

3.2.6 Moving PCA

The monitoring performance of conventional PCA can be improved upon through the application of moving PCA (MPCA) and its associated monitoring indices (Kano et al., 2001), resulting from PCA being applied on-line by moving a time-window. In MPCA changes of operating condition, characterised by a change in the correlation structure among the process variables and not changes in the scores of predefined PCs, are detected through changes in the direction of each PC or changes in the subspace spanned by several PCs.

Changes in the direction of each PC is monitored using the index A_i. The change of the i th PC is evaluated:

$$A_i(j) = 1 - \left| l_i(j)^T l_{i0} \right| \tag{3-23}$$

where $l_i(j)$ denotes the i th PC at step j, and l_{i0} denotes the reference of i th PC (both $l_i(j)$ and l_{t0} being unit vectors). When the i th PC representing the current operating condition is equivalent to that of the reference, A_i becomes zero. When $l_i(j)$ is orthogonal to l_{i0}, A_i becomes one. MPCA can detect changes of correlation among process variables, which conventional MSPC with the Hotelling's T^2 and Q statistics may find difficult to detect. When variances of several PCs are similar to each other, the index A_i does not function well. This results from the directions of PCs changing abruptly while the correlation among the process variables remains unchanged. This problem is addressed by monitoring a change of subspace spanned by several PCs with similar variances instead of a change of each PC:

$$A_{1-m} = \sqrt{1 - \lambda_{\min}} \tag{3-24}$$

where $\lambda_{\min}$ is the minimum eigenvalue of the matrix $W^T P W$ (derived for a change of subspace) and A_{1-m} is the index for monitoring the change of subspace spanned by the first m PCs. The index A_{1-m} enables MPCA to function successfully even when variances of several PCs are similar to each other.

The procedure for implementing MPCA is as follows:

1. Acquire NOC time-series data, auto-scaling each column (variable) of the data matrix.
2. Apply PCA to the data matrix, and define a reference PC, l_{i0}.
3. Determine the size (steps) of time-window, w. Generate data sets with w samples from the NOC data by moving the time-window. Apply PCA to the data sets, and calculate PCs, l_i.
4. Calculate the index A_i and/or A_{1-m}, and determine the control limits.

Stellenbosch University http://scholar.sun.ac.za

5. For on-line monitoring update the data matrix representing the current operating condition by moving the time-window step by step.
6. Scale the updated data matrix with the mean and the variance obtained at step 1.
7. Apply PCA to the $W \times P$ data matrix.
8. Calculate the index A_i and/or A_{1-m}.
9. Repeat from step 5. If the index is outside the control limit, the process is judged to be out of control.

MPCA has been shown to have considerably better reliability when compared to conventional PCA. It is able to detect changes in operating conditions even when the deterministic changes in the variables being monitored are not significant and the variance is not increased. However, MPCA has a smoothing effect, caused by the use of a time-window, and therefore suffers from a delay in detecting events and returns to the normal operating conditions. This makes the selection of an appropriate size of time-window crucial for the effective functioning of MPCA.

3.2.7 Process data dissimilarity

Although many successful applications have shown the practicability of multivariate SPC based on conventional PCA and Hotelling's T^2 and the Q statistic, it does not always function well due to it not being able to detect correlation changes among process variables as long as both Hotelling's T^2 and the Q statistic are within their respective control limits. The process data dissimilarity index, DISSIM, (Kano et al., 2002) was introduced to improve process monitoring performance. It is based on the idea that a change in operating condition can be detected by monitoring a distribution of time-series data, which reflects the corresponding operating condition.

The differences between distributions of data sets are evaluated using a classification method based on the Karhunen-Loeve expansion, and it is mathematically equivalent to PCA. Given 2 data sets, X_i, the covariance matrix of the mixture of both data sets is estimated:

$$R = \frac{n_1 - 1}{n - 1} R_1 + \frac{n_2 - 1}{n - 1} R_2 \qquad (3\text{-}25)$$

where R_i is the covariance matrices of the individual data sets, n_i the number of samples in the individual data sets and n the total number of samples. Eigenvalue decomposition is performed and the transformation matrix obtained:

$$P = l\Lambda^{-1/2} \qquad (3\text{-}26)$$

where l is the eigenvectors and λ the eigenvalues. Using the transformation matrix, the individual matrices are then transformed:

$$Y_i = \sqrt{\frac{n_i - 1}{n-1}} X_i P \tag{3-27}$$

Through calculation of the covariance matrices of these transformed data matrices and their subsequent eigenvalue decomposition, it is found that the most important correlation for data set 1 is equivalent to the least important correlation for data set 2, and vice versa. When data sets are quite similar to each other, the eigenvalues $\lambda_j^{(i)}$ (where i refers to the data set number and j refers to the eigenvalue number) must be near 0.5. On the other hand, when data sets are quite different from each other, the largest and the smallest eigenvalues should be near one and zero, respectively. The dissimilarity index can then be defined as:

$$D = \frac{4}{p} \sum_{j=1}^{p} \left(\lambda_j - 0.5\right)^2 \tag{3-28}$$

where p is the number of variables. The dissimilarity index changes between zero, for similar data sets, and one, for dissimilar data sets.

The procedure for implementing the process data dissimilarity index is as follows:

1. Acquire NOC time-series data, auto-scaling each column (variable) of the data matrix.
2. Determine the size (steps) of the time window, w. Generate data sets with w samples from the data by moving the time window. Select a reference data set. It is important that the size of the reference data set not exceed that of the time window. One could also randomly select w samples from the data matrix.
3. Calculate the dissimilarity index, D, and the control limits.
4. For on-line monitoring update the data matrix representing the current operating condition by moving the time-window step by step.
5. Scale the updated data matrix with the mean and the variance obtained at step 1.
6. Calculate the dissimilarity index, D.
7. Repeat from step 4. If the index falls outside the control limit, the process is judged to be out of control.

DISSIM have been shown to have considerably better reliability when compared to conventional PCA. However, similar to MPCA, DISSIM has a smoothing effect, caused by the use of a time-window, and therefore also suffers from a delay in detecting events and returns to the normal operating conditions. This makes the selection of an appropriate size of time-window crucial for the effective functioning of

Stellenbosch University http://scholar.sun.ac.za

DISSIM. It should also be noted that DISSIM is very sensitive to changes in the correlation structure of the data being investigated.

3.3 Dynamic multivariate statistical process monitoring

Although conventional PCA takes the multivariate nature of the data into account, the dynamic behaviour of the process is not being considered. The application of conventional PCA to continuous process data may, therefore, be inadequate. This shortfall results from the time dependency of a dynamic system when considering the variable histories. For dynamic processes, both the correlation among variables and the autocorrelation of each variable should be considered. This requires dynamic models.

3.3.1 Dynamic PCA

Ku et al. (1995) extended conventional PCA through a well-known time lagged variable method to include dynamic behaviour in the PCA model. When considering a system with dynamics, the current values of any variable will depend on past values of that variable. Therefore, the linear relations between $X(j)$ and $X(j-1)$, at least, need to be identified. If all dynamic relationships are first order systems, the relation will be the noise subspace of the following equation:

$$[X(j)X(j-1)]b = 0 \tag{3-29}$$

In a more general case,

$$X_{\mathrm{T}}(w)b = 0 \tag{3-30}$$

where

$$\begin{aligned} X_{\mathrm{T}}(w) &= [X(j)X(j-1)\ldots X(j-w)] \\ &= \begin{bmatrix} x^T(1) & x^T(0) & \ldots & x^T(1-w) \\ x^T(2) & x^T(1) & \ldots & x^T(2-w) \\ \vdots & \vdots & \ddots & \vdots \\ x^T(n) & x^T(n-1) & \ldots & x^T(n-w) \end{bmatrix} \end{aligned} \tag{3-31}$$

Dynamic PCA (DPCA) is essentially the same as the conventional PCA approach except that the data matrix is composed of time lagged duplicate vectors. The proper choice of the number of time lags, w, will ensure that both the static and dynamic relations should appear in the noise subspace with small singular values. The number of time lags is usually 1 or 2, indicating the order of the dynamic system. For non-linear systems, increasing w could result in a better linear approximation of the actual non-linear relationships.

Stellenbosch University http://scholar.sun.ac.za

The procedure for determining the proper number of PCs to retain and the order of the system is as follows:

1. Set $w = 0$.
2. Form the data matrix $\left[X(j)X(j-1)\cdots X(j-w)\right]$.
3. Perform PCA and calculate all the principal scores.
4. Set $i = p \times (w+1)$ and $r(w) = 0$.
5. Determine if the i th component represents a linear relation. If yes proceed, if no go to step 7.
6. Set $i = i-1$ and $r(w) = r(w)+1$, repeat step 5.
7. Calculate the number of new relationships

$$r_{new} = r(w) - \sum_{i=0}^{w-1} (w-i+1) r_{new}(i) \qquad (3\text{-}32)$$

8. If $r_{new}(w) \leq 0$, stop, otherwise proceed.
9. Set $w = w+1$, go to step 2.

It is suggested that one use parallel analysis and the cross-correlation plots of the scores to determine the number of the PCs. The number of linear relations, r, will then be the total number of the variables minus the number of the PCs. It is important to check if the remaining scores are completely independent using the auto- and cross-correlation plots. All the static monitoring indices can be used with the DPCA models. It has been shown that DPCA performs better than conventional PCA in detecting the occurrence of small disturbances in a dynamic process.

Tsung (2000) extended DPCA and proposed a modified and simplified procedure. In short, data for modelling is not only collected under NOC, but also under various known types of out-of-control conditions (OOCC) and the number of PCs to retain and the order of the system is not iteratively determined, but directly from the analytical model. The procedure proposed to monitor and diagnose an automatic controlled process is as follows:

1. Estimate an analytical model of the process, using this model to determine the necessary number of time lags.
2. Collect both NOC and special OOCC data. OOCC data is required to construct an OOCC database for diagnosis.
3. Build a DPCA model with time lags as identified in step 1 for each of the specified OOCC, thus constructing the OOCC database.
4. Monitor the process using Hotelling's T^2 and the Q statistic. If an out-of-control signal is detected, proceed to step 5 to diagnose the process.
5. Observations are mapped to the OOCC database for root cause isolation and identification. After identifying the most possible OOCC, confirm it by checking its Q statistics. It is only concluded that the out-of-control process is due to the OOCC if the Q chart agrees. Should the case

Stellenbosch University http://scholar.sun.ac.za

observations of the out-of-control process not be consistent with the classified OOCC model, the Q chart will indicate the presence of a new OOCC. Once this root cause is isolated and identified, the new model can be added to the OOCC database for future use.

Lin et al. (2000) combined the strength of DPCA, for extracting time-dependent relationships in measurements by augmenting the data matrix by time lagged variables, and non-linear PCA using neural networks, for extracting non-linear relationships among process variables. The procedure proposed for the implementation of non-linear DPCA (ND-PCA) for process monitoring and diagnosis is as follows:

1. Obtain a NOC data matrix.
2. Use DPCA to filter the raw data.
3. Use the DPCA scores to develop a score predictive model, dividing the dynamic measurement space into two subspaces: a PC subspace and a residual subspace.
4. For every event scenario, the developed DPCA model is used to rectify the data matrix and the projection of every event scenario in PC subspace is obtained. The projections of each event in the PC subspace is then classified using a sigmoid basis function feed forward neural network. The Levenberg-Marquardt algorithm used to train this neural network was improved upon through the use of a genetic algorithm to guarantee the global optimisation of the network and a low training error.
5. Monitor the process using Hotelling's T^2 and the Q statistic.
6. If an out-of-control signal is detected the sample vector's projection in PC subspace is fed into the neural network to implement event identification.

3.3.2 Summed-scores PCA

Wachs and Lewin (1999) addressed the shortfall of the conventional PCA methodology not accounting for time-dependent relationships among process variables by recursively summing the last w PCA scores, referred to as summed-scores PCA (SSPCA), and using these summed-scores to construct the descriptive statistics for process monitoring. This approach extends the standard univariate moving-average techniques, to the multidimensional space of scores, obtained from applying PCA. Additionally, for processes characterised by time-varying trajectories, delay-adjusted PCA (DAPCA) is used to account for the dynamic delays between variables by applying relative shifts between the inputs and outputs. The most appropriate time delays between input and output trajectories are determined by shifting input variables backward until their correlation with the outputs are maximized. DAPCA has similarities with dynamic PCA (Ku et al., 1995) in which the dynamic behaviour of the process is treated but not its delays.

For SSPCA, the w-summed construct is computed as follows:

$$\bar{t}_{1,j} = \sum_{k=j-w+1}^{j} t_{1,k} \quad , \quad j = w,\ldots,n \tag{3-33}$$

Stellenbosch University http://scholar.sun.ac.za

where n is the number of samples, i is the number of the score being summed and w the time window over which the scores are summed. It can be shown that $\overline{T_1}$, the first vector of the summed-scores scores matrix, has a normal distribution with zero mean and standard deviation $\overline{\sigma}$. If the NOC boundaries are at $\pm 3\overline{\sigma}$ from the origin, 99.7% of NOC data will lie within the NOC limits. Therefore using SSPCA, the NOC limits are at $3\times\sqrt{w}\times\sigma$ compared with 3σ limits for conventional PCA. This increases the resolution between the NOC and the failure point and could often lead to faster event detection. By increasing w, the resolution will be increased, although at the price of delayed reaction.

It was found that the appropriate number of samples to be summed, w, needs to be large enough to enable small shifts to be detected, but not larger than necessary to reduce the diagnosis reaction time to a minimum. SSPCA proved to be especially successful in detecting small shifts where conventional PCA was found to be inadequate.

The algorithm for DAPCA is as follows:

1. Form the $n\times p$ raw data matrix, X, consisting of n rows (observations) and p columns (variables), the first m of which are inputs and the subsequent $p-m$ are outputs.
2. Define the maximum reasonable process delay, $d_{\max}$, in terms of samples.
3. For each input variable, i:
    a. Form the $n\times(p-m+1)$ data matrix, X_i, from the i th input and all $p-m$ outputs.
    b. Select the optimal backward shift in the i th input, d_i, in the range $0\le d_i\le d_{\max}$, such that the determinant of the correlation matrix computed from X_i is minimised.
    c. Store the optimal shift for the i th input, d_i.
4. Adjust the original data set such that each input vector is shifted by its optimal delay.
5. Perform PCA on the modified data set, and adjust new samples by the optimal shifts computed in step 3.

The DAPCA algorithm assumes the inputs to be independent of one another and the output variables to be correlated among themselves with no delays present. It is therefore not possible to guarantee that the proposed algorithm always identifies the optimal shifts that find the global minimum for the object function.

3.3.3 Multiscale PCA

In order to model data containing contributions from events whose behaviour changes over time and frequency, Bakshi (1998) suggested the use of multiscale PCA (MSPCA). MSPCA combines the ability of PCA to decorrelate the variables by extracting linear relationships with that of wavelet analysis to extract deterministic features and approximately decorrelate auto-correlated measurements.

Stellenbosch University http://scholar.sun.ac.za

The advantage of wavelets lies in the fact that this family of basis functions are localised in both time and frequency:

$$\psi_{sv}(k) = \frac{1}{\sqrt{s}}\psi\left(\frac{k-v}{s}\right) \tag{3-34}$$

where $\psi(k)$ is the mother wavelet, s represents the dilation parameter (determining the location of the wavelet in the frequency domain as well as the scale or extent of the time-frequency localisation) and v represents the translation parameter (determining the location of the wavelet in the time domain). By projection on the corresponding wavelet basis function, any signal may be decomposed into its contributions in different regions of the time-frequency space. The number of coefficients for both the wavelet and scaling functions decreases dyadically at coarser scales. This is a direct result of the dyadic discretisation of the dilation and translation parameters. Convolution with a filter I and J represents projections on the scaling function and wavelet respectively. The coefficients at different scales can be obtained:

$$a_m = Ia_{m-1} \quad , \quad d_m = Ja_{m-1} \tag{3-35}$$

where d_m is the vector of wavelet coefficients at scale m and a_m is the vector of scaling function coefficients.

The algorithm for MSPCA is as follows:

1. For each column in the data matrix
   Compute the wavelet decomposition.
   End
2. For each scale
   Compute the covariance matrix of the wavelet coefficients.
   Compute the PCA eigenvectors and scores of the wavelet coefficients.
   Select the appropriate number of principal components to retain.
   Select the wavelet coefficients larger than an appropriate threshold.
   End
3. For all the scales together
   Compute the PCA by including all the scales with significant events.
   Reconstruct the approximate data matrix from the selected and thresholded scores at each scale.
   End

The algorithm can be applied on-line by decomposing the data in a moving window of dyadic length with the most recent sample included in the window. Spurious features created by sudden changes can be

Stellenbosch University http://scholar.sun.ac.za

decreased by averaging the signal reconstructed from each moving window. This will also improve the accuracy and smoothness of the extracted features. MSPCA has not only been found to outperform conventional PCA but also integrates the task of feature extraction and process monitoring.

Fourie and de Vaal (2000) developed a non-linear multiscale principal component analysis (NLMSPCA) methodology to detect deterministic changes and extract those features that represent abnormal operation. NLMSPCA combines the strength of non-linear PCA via an input-training neural network, for extracting both linear and non-linear relationships from measurements, and wavelet analysis, for extracting deterministic features and approximately decorrelating auto-correlated measurements. In addition, non-parametric control limits are calculated for the control charts used to identify the occurrence of out-of-control situations. The procedure proposed for the implementation of NLMSPCA is as follows:

1. Pre-screen data to identify and handle outliers, missing data, etc.
2. Apply multiresolution analysis based on wavelets, decompose each variable into its contributions in different regions of the time-frequency space.
3. Apply level dependent thresholding to the wavelet coefficients at each scale, selecting a smaller subset of wavelet coefficients, also separating the stochastic and deterministic components of the signal.
4. Details and approximations are reconstructed in the time domain from both the thresholded and non-thresholded wavelet coefficients. Thresholded details and approximations containing significant contributions are retained in a data set 1, whereas non-thresholded details and approximations are retained in a data set 2. Steps 5 to 10 are applied to both data set 1 and data set 2.
5. Both linear and non-linear PCA is performed independently on the details and approximations at each scale.
6. Non-linear PCA is performed based on the input-training neural network approach.
7. For on-line application, a mapping model is developed between the process observations and the non-linear principal scores. A feed-forward neural network is used to combine the mapping network and the input-training neural network.
8. Bivariate non-linear PC score plots and SPE plots are derived for performance monitoring. Control limits are determined based on non-parametric density estimation using kernel estimation.
9. An out-of-control situation is detected when the current scores and residuals violate the control limits.
10. Define a differential contribution plot, describing the difference between the contributions of the process variables to their non-linear scores.

Kano et al. (2002) extended MSPCA by incorporating the moving PCA and DISSIM with their associated statistics into the MSPCA methodology. The monitoring procedure of multiscale MPCA or multiscale DISSIM is the same as that of MSPCA, except that wherever conventional PCA was applied MPCA or DISSIM is now applied, and the T^2 and Q indices are replaced by A_i or D.

Stellenbosch University http://scholar.sun.ac.za

3.4 Non-linear multivariate statistical process monitoring

Although great success has been achieved with the application of linear dimensionality reduction techniques such as linear PCA for multivariate SPM, many real processes are non-linear. Non-linear extensions of linear PCA are therefore one way of enhancing multivariate SPM (Dong and McAvoy, 1996; Zhang et al., 1997; Jia et al., 1998). With non-linear PCA, the first two principal components normally also explain more of the variance in the data than is possible through linear PCA, enhancing the graphical interpretation of the results.

3.4.1 Conventional PCA and Neural networks

Jia et al. (1998) combined conventional PCA and neural networks, developing a non-linear PCA approach consisting of three main steps. During the first step linear PCA is applied to the data, extracting linear information from the data with sufficient data variance being retained in the transformed data to ensure that non-linear correlations are not removed from the model. During the second step the PCA scores are scaled to unit variance, compressing the linear structure of the data and so allowing the recovery of the non-linear structure in the transformed data. The last step of the approach entails capturing the latent non-linear structure of the transformed data through modelling using a multilayer perceptron (MLP) neural network. This allows both linear and non-linear information to be captured in the final non-linear principal component scores.

Following this work, Chen and Liao (2002) also combined conventional PCA and neural networks by developing a technique called neural network PCA (NNPCA), applicable to both linear and non-linear systems. The methodology consists essentially of two core stages: residual generation and residual evaluation. Residuals are generated by comparing the actual behaviour of the process to be supervised with that of a nominal event-free neural network model driven by the same observations. The residuals derived from the difference between the actual process and the neural network predictions are subsequently evaluated using PCA. It is expected that the residuals be closed to zero under the NOC. Under abnormal operating conditions, the zero point of the residual variable would drift away. Comparing the residuals with a decision function or predefined threshold from the NOC statistical analysis is done in order to determine if the new process behaviour can be classed as in-control or out-of-control. Multivariable control charts are used to monitor the residuals due to the correlations between the residual variables. Therefore, the neural network acts as a non-linear dynamic operator used to remove the non-linear and dynamic characteristics, whereas PCA is applied to generate simple monitoring charts.

In order to determine control limit thresholds for process monitoring, the residual data for the NOC data is computed from the predicted neural network and put in a 2-dimensional residual matrix $E\ (n \times p)$:

Stellenbosch University http://scholar.sun.ac.za

$$E = \left[e_j^T \;\; e_{j+1}^T \;\; \dots \;\; e_n^T \right]^T = \begin{bmatrix} e_1(j) & e_2(j) & \dots & e_p(j) \\ e_1(j+1) & e_2(j+1) & \dots & e_p(j+1) \\ \vdots & \vdots & \dots & \vdots \\ e_1(n) & e_2(n) & & e_p(n) \end{bmatrix} \quad (3\text{-}36)$$

with n samples and p prediction residual variables. This residual matrix can be decomposed into linear PCs. Using the derived PCs and scores, Hotelling's T^2 and the Q statistics can be used to determine thresholds for the control limits.

NNPCA has been found to be more effective than the conventional PCA or DPCA, especially when the process exhibits dynamic and non-linear behaviour. The sensitivity and robustness of NNPCA for process monitoring can also be improved by integrating CUSUM or EWMA directly into NNPCA. By doing so, EWMA can be used first to determine the predicted residuals from the neural net model and the moving average variables, and then PCA can be used to extract the correlation between the moving average variables.

3.4.2 Auto-associative neural networks

Instead of combining linear PCA with neural networks, Kramer (1991) used an auto-associative neural network for deriving non-linear principal components from data. This is accomplished through training of a feedforward neural network, reproducing model inputs at the model output layer. The auto-associative neural network consists of an input layer (linear or non-linear), three hidden layers and an output layer (linear or non-linear). The hidden layers consist of a non-linear mapping/enconding layer, a linear or non-linear bottleneck layer and a non-linear demapping/decoding layer. The non-linear nodes in the mapping and demapping layers are required to ensure the generation of non-linear combinations of the model inputs with the bottleneck layer containing fewer nodes than either the input or output layer, forcing the network to develop a compact representation of the input data.

There are two methods for calculating non-linear principal components using an auto-associative neural network. For the first method the size of the bottleneck layer is determined by the total number of non-linear principal components required. Subsequently, a single neural network is trained, with the output of the bottleneck layer representing the required non-linear principal components. The alternative method, auto-associative neural network principal component analysis (AANNPCA), requires multiple neural networks, each containing a single node in the bottleneck layer, to be sequentially trained (Scholz et al., 2008). Starting with the first neural network, only the primary non-linear principal component is extracted by the bottleneck layer. The residual from the first network becomes the input to the second neural network. The second neural network is subsequently trained and the whole process repeated until the required number of non-linear principal component has been extracted. Sequential training of single

Stellenbosch University http://scholar.sun.ac.za

node bottleneck layer networks not only allows for rescaling of the residuals between training steps but also forces the bottleneck nodes to model separate factors in the data.

Using auto-associative neural networks for extracting non-linear principal components have proved to be more effective than conventional linear PCA in describing and reducing data. When tested on batch reaction data, non-linear PCA significantly outperformed conventional linear PCA.

3.4.3 Principal curves

Principal curves are a non-linear dimensionality reduction technique. For an m-dimensional data set, the smooth, 1-dimensional curve passing through its middle is termed a principal curve (Dong and McAvoy, 1996). The structure of the data determines the shape of the principal curve, which in turn provides a non-linear summary of the data (Figure 9). If the principal curve is a straight line, it is a linear principal component. The principle curve can be represented by a vector $f(c)$ of coordinate functions, of a single variable c. The variable c, often the arc length along the curve, parameterises the curve and provides an ordering along it. Given $x \in R^p$, a continuous random vector with a known distribution H, the curve f is a principal curve of H if

$$\mathrm{E}\left[x \middle| c_f(x) = c\right] = f(c) \tag{3-37}$$

where c_f is defined as a projection index of $R^p \rightarrow R^t$. The value of c for which $f(c)$ is closest to x is defined as the projection index $c_f(x)$ of x. If several such values exist, the largest one is used.

For estimation of the principal curve, the expectation operator, E, requires that the distribution of x be known. However, real data is often multivariate and finite, $x \in R^p$, with unknown distribution. $\mathrm{E}\left[X \middle| c_f^i(X) = c\right]$ can be estimated by means of scatter plot smoothing and locally weighted regression. The algorithm used is iterative and consists of two steps:

1. Project the data onto $\hat{f}_i$, the current estimate of the principal curve, and assign a parameter value c_i (the arc length measured from the starting point on the curve) to each data point. The data is then ordered according to the parameter values, defining the neighborhoods for use in the next step.
2. Use locally weighted regression and smoothing scatter plots to estimate $\mathrm{E}\left[X \middle| c_f^i(X) = c\right]$. Calculate the Euclidean distance between the data set and the estimated principal curve. If the relative change exceeds some threshold, repeat from step 1.

Stellenbosch University http://scholar.sun.ac.za

For the application of locally weighted regression, using only the data points in a neighborhood and not all for regression, span is an important parameter to set. The span, the fraction of the data points that are considered to be in the neighborhood, controls the size of the neighborhood. Increasing the span increases the smoothness of the fit. Decreasing the span interpolates the data points. Unfortunately, the principal curve algorithm (Dong and McAvoy, 1996) is not robust for outliers, which easily influence the estimation of the conditional expectation. The principal curve algorithm also does not produce a non-linear PC model in the sense of a PC loading, and only calculates an associated score and corrected data point for each input data point. The principal curve algorithm, therefore, only gives a generalization of the first linear principal component. However, using neural networks the scores can be modeled to the corrected data points, effectively developing a non-linear PC model (NLPCA) when combined with the principal curve algorithm.

By orthogonally projecting the data in the p-dimensional data set X onto the principal curve, a countable ordering of the data is obtained. Each data point now has a corresponding length along the curve, its length defined as the non-linear principal component score of that data point. For the p-dimensional data set X having n data points, its non-linear principal scores form an $n \times 1$ column vector. With the score vector represented by t_i, X can be modeled as

$$X = f_1(t_1) + E_1 \tag{3-38}$$

where E_1 is the residual. For each additional non-linear PC required, the calculation is repeated for the residual data until no information is contained in the residual data. The residual can be calculated as

$$E_{l-1} = f_1(t_1) + E_1 \tag{3-39}$$

or more generally

$$X = \Gamma(\mathrm{T}) + E \tag{3-40}$$

where E is the residual and $\mathrm{T} = [t_1, t_2, \ldots, t_l]$ is the non-linear principal score matrix.

$\Gamma(\bullet)$, the non-linear function, is defined as the non-linear principal loading function with the variance explained being used to choose the number of non-linear PCs. Neural networks (good universal approximators) such as the auto-associative neural network can effectively be used to model $\Gamma(\bullet)$ when performing NLPCA. Using principal curves and an auto-associative neural network for NLPCA allows not only the non-linear principal scores and corrected data set to be obtained, but also the non-linear principal loading functions.

Stellenbosch University http://scholar.sun.ac.za

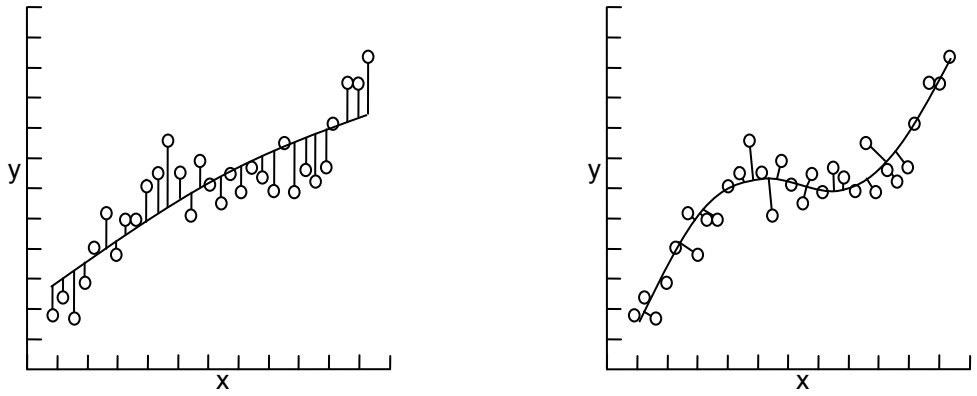

Figure 9: **An illustrative example showing non-linear regression minimising the sum of squared deviations using a non-linear function (a), and non-linear principal curves using a smooth curve to minimise the sum of squared orthogonal deviations (b). (adapted from Dong and McAvoy, 1996)**

For process monitoring, non-linear PCs are extracted from the process data and an SPE -chart or score-chart is used for monitoring. It has been shown that NLPCA can effectively capture the non-linear relationship in process variables, outperforming linear PCA, and SPE - and score-charts can be used to detect process abnormalities.

Zhang et al. (1997) extended this method of process monitoring using NLPCA by also monitoring accumulated scores, thus accounting for process dynamics. Accumulated scores are defined as:

$$A(j) = \sum_{i=1}^{j} \left(t(i) - \bar{t} \right) \tag{3-41}$$

where the non-linear score for the i th observation is $t(i)$, the mean of the nominal non-linear score is $\bar{t}$, and the accumulated non-linear score till time n is $A(j)$. During normal operation $A(j)$ will remain around the origin, however, once an event occurs $A(j)$ will move away from the origin with different events causing movement into different directions.

3.4.4 Extreme learning machine PCA

Extreme learning machine (ELM) provides an opportunity with which to improve the neural network PCA algorithm proposed by Chen and Liao (2002). ELM is a learning algorithm for single hidden layer feedforward neural networks (SLFNs) where the input weights are chosen randomly and the output weights determined analytically (Huang et al., 2006). Whereas traditional learning algorithms employed by feedforward neural networks are limited in their learning speed, mainly due to the slow gradient descent-based nature of the algorithms and the iterative parameter estimation, ELM can not only achieve

extremely fast learning speed but also tends to provide good generalization performance (thanks to the learning algorithm not only typically reaching the smallest training error but also the smallest norm of weights). Furthermore, for conventional feedforward neural networks there exists a dependency between the different layers of weight and bias parameters which requires all parameters to be tuned. For SLFNs with N hidden nodes it has been shown that with randomly chosen input weights and hidden layer biases exactly N distinct observations can be learned (Huang, 2003), limiting the aforementioned dependency and significantly reducing the number of parameters that needs to be estimated. This allows such a SLFN to be considered as a linear system with the output weights being analytically determined through simple generalized inverse operation of the hidden layer output matrices (Huang et al., 2006). However, the random selection of the input weights and biases can easily cause the hidden layer output matrix to be not full column rank (Wang et al., 2011). In turn, this sometimes leads to the linear system used for training the output weights being unsolvable, lowering the ELM predicting accuracy and in so doing, lowering the effectiveness of the ELM.

Given a training data set consisting of N arbitrary distinct samples (x_j, y_j), a hidden node number $\tilde{N}$, and an activation function $g(x)$, the ELM algorithm can be defined as (Huang et al., 2006):

1. Randomly assign the input weights z_j and biases b_j, $j = 1, \ldots, \tilde{N}$.
2. Calculate the hidden layer output matrix P.
3. Calculate the output weights $\beta = P^H Y$ where $Y = [y_1, \ldots, y_N]^T$

Considering the above, using an ELM trained SLFN will not only result in a simpler form of the neural network PCA algorithm, but also in a faster, more robust and potentially more reliable version, making it ideal for use when analyzing very large data sets.

The proposed ELM PCA methodology is therefore almost identical to the NNPCA methodology, the exception being the conventional feedforward neural network being replaced by an ELM trained SLFN. As with the NNPCA methodology, the ELM PCA methodology can be considered to consist of two core stages: residual generation and residual evaluation. Residuals are generated by comparing the actual behaviour of the process to be supervised with that of a nominal event-free ELM trained SLFN model driven by the same observations. The residuals derived from the difference between the actual process and the SLFN predictions are subsequently evaluated using PCA. The residuals are expected to be close to zero under the NOC. Under abnormal operating conditions, the zero point of the residual variable would drift away. Comparing the residuals with a decision function or predefined threshold from the NOC, statistical analysis is done in order to determine if the new process behaviour can be classed as in-control or out-of-control. Multivariable control charts are used to monitor the residuals due to the correlations between the residual variables. Therefore, the SLFN acts as a non-linear dynamic operator used to remove the non-linear and dynamic characteristics, whereas PCA is applied to generate simple monitoring charts.

Stellenbosch University http://scholar.sun.ac.za

3.5 Technique evaluation case studies

Various benchmark models and data sets exist with which to evaluate process monitoring techniques. These range from simple multivariate time series data to complex dynamic models and real world case studies of process data. Some data present easily identifiable events in individual time series data, such as a spike in a signal, whereas other data present slow complex multivariate change in many process signals simultaneously, such as a change in the relationship between two variables. Each model and data set is unique in that it can be used to evaluate specific characteristics of the techniques applied, allowing the performance of the various process monitoring techniques to be accurately evaluated.

For the purpose of evaluating techniques and methodologies prior to application on the concentrator case study the following data sets and models have been selected for validation purposes:

- Simple multivariate time series data – described in detail in section 3.5.1
- Simple multivariate process (Ku et al., 1995) – described in detail in section 3.5.2
- Tennessee Eastman process (Downs and Vogel, 1993) – described in detail in section 3.5.3

This evaluation also allows for an appreciation of the techniques and their applicability to be obtained.

3.5.1 Simple multivariate time series data

Multivariate time series data $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}$ consisting of 1000 samples for 4 variables are generated having a specified mean of $\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$ and a covariance of $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0.5 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0.5 & 0.5 & 1 \end{bmatrix}$ with random, Gaussian noise with zero mean and a standard deviation of $\begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$.

The set of simulated abnormal conditions introduced at sample index 101 of 1000 consists of (Figure 10):

MTS0 - Normal operating data

- A step change (off-set) in the mean value of one of the variables:

  MTS1 - A mean shift in x_1 from 0 to 0.5

  MTS2 - A mean shift in x_1 from 0 to 1

  MTS3 - A mean shift in x_3 from 0 to 0.5

  MTS4 - A mean shift in x_3 from 0 to 1

- A ramp change (drift) in the mean value of one of the variables:

  MTS5 - A mean shift in x_1 at a rate of 0.005 per sample

  MTS6 - A mean shift in x_1 at a rate of 0.01 per sample

  MTS7 - A mean shift in x_3 at a rate of 0.005 per sample

Stellenbosch University http://scholar.sun.ac.za

MTS8 - A mean shift in x_3 at a rate of 0.01 per sample

- A spike with a peak value of 2 in one of the variables:

  MTS9 - A spike in x_1 having a duration of 1 sample

  MTS10 - A spike in x_1 having a duration of 10 samples

  MTS11 - A spike in x_3 having a duration of 1 sample

  MTS12 - A spike in x_3 having a duration of 10 samples

- A step change (off-set) in the standard deviation of one of the variables:

  MTS13 - A standard deviation shift in x_1 from 0.1 to 0.5

  MTS14 - A standard deviation shift in x_1 from 0.1 to 1

  MTS15 - A standard deviation shift in x_3 from 0.1 to 0.5

  MTS16 - A standard deviation shift in x_3 from 0.1 to 1

- A step change (off-set) in the covariance of one of the variables:

  MTS17 - A covariance shift to $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0.5 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0.5 & 0.5 & 1 \end{bmatrix}$

  MTS18 - A covariance shift to $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0.5 & 0.5 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0.5 & 0.5 & 1 \end{bmatrix}$

  MTS19 - A covariance shift to $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.8 & 0.5 \\ 0 & 0.8 & 1 & 0.8 \\ 0 & 0.5 & 0.8 & 1 \end{bmatrix}$

- A step change (off-set) in the function type of one of the variable:

  MTS20 - A change in the function type of x_1 to a sine wave with an amplitude of 2

  MTS21 - A change in the function type of x_1 to a sine wave with an amplitude of 5

  MTS22 - A change in the function type of x_3 to a sine wave with an amplitude of 2

  MTS23 - A change in the function type of x_3 to a sine wave with an amplitude of 5

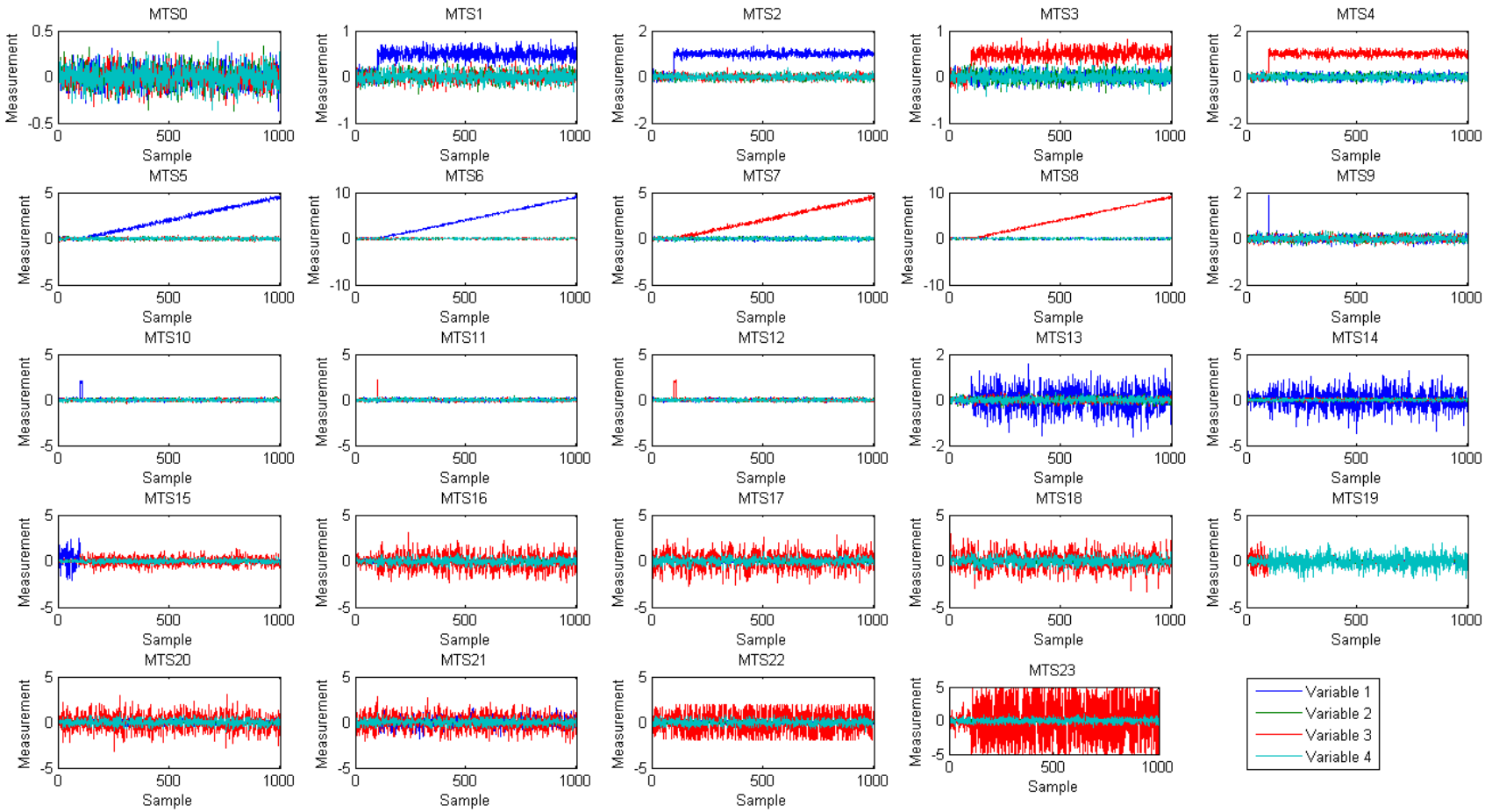

Figure 10: **Simple multivariate time series data: simulated conditions**

Stellenbosch University http://scholar.sun.ac.za

3.5.2 Simple multivariate process

For validation purposes Ku et al. (1995) defined the following simple 2x2 multivariate process:

$$c(j) = C_{SMP}.c(j-1) + D_{SMP}.x(j-1) \tag{3-42}$$

$$y(j) = c(j) + e_1(j) \tag{3-43}$$

where $C_{SMP} = \begin{bmatrix} 0.118 & -0.191 \\ 0.847 & 0.264 \end{bmatrix}$, $D_{SMP} = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$ and x is the correlated input:

$$x(j) = A_{SMP}.x(j-1) + B_{SMP}.e_2(j-1) \tag{3-44}$$

where $A_{SMP} = \begin{bmatrix} 0.811 & -0.226 \\ 0.477 & 0.415 \end{bmatrix}$, $B_{SMP} = \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix}$, the input e_2 is random, Gaussian noise with zero mean and unit variance and the output y is equal to c plus the random noise, $e_1(j)$, with zero mean and variance 0.1. Both input, x, and output, y, are measured but c and e_2 are not.

For monitoring purposes the following 4 variables are used (each variable simulated to contain 1000 samples):

1. Model outputs y (2 variables).
2. Correlated model inputs x (2 variables).

The set of simulated abnormal conditions introduced at sample index 101 of 1000 consists of (Figure 11):

SMP0 - Normal operating data

SMP1 - A step change in the mean of $e_2(1)$ from 0 to 0.5.

SMP2 - A step change in the mean of $e_2(1)$ from 0 to 1.

SMP3 - A step change in the mean of $e_2(1)$ from 0 to 1.5.

SMP4 - A step change in the mean of $e_2(1)$ from 0 to 2.

SMP5 - A step change in the mean of $e_2(1)$ from 0 to 3.

SMP6 - A step change in the parameters of $D_{SMP}(2,1)$ from 3 to 2.5.

SMP7 - A step change in the parameters of $D_{SMP}(2,1)$ from 3 to 2.

SMP8 - A step change in the parameters of $D_{SMP}(2,1)$ from 3 to 1.

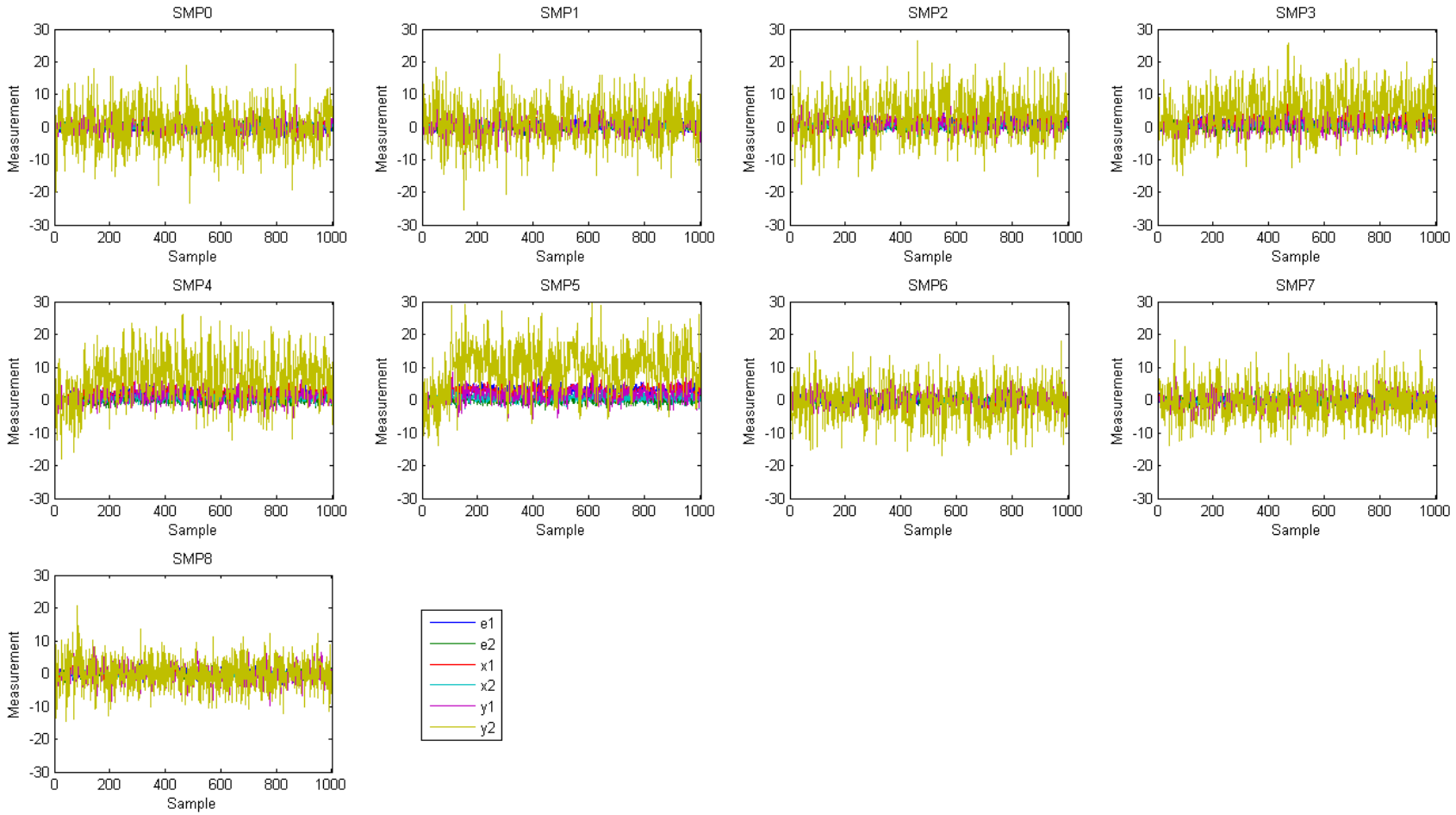

Figure 11: **Simple multivariate process: simulated conditions**

Stellenbosch University http://scholar.sun.ac.za

3.5.3 Tennessee Eastman process

A more difficult process to monitor is the chemical plant simulation developed by Downs and Vogel (1993) of the Eastman Chemical Company. The process consists of a reactor/separator/recycle arrangement (Figure 12). Products are produced by two simultaneous gas-liquid exothermic reactions of the following form:

$$A_{TE}(gas)+C_{TE}(gas)+D_{TE}(gas)\rightarrow G_{TE}(liquid) \quad (3\text{-}45)$$

$$A_{TE}(gas)+C_{TE}(gas)+E_{TE}(gas)\rightarrow H_{TE}(liquid) \quad (3\text{-}46)$$

A by-product is generated by two additional exothermic reactions. The process has 12 manipulated variables, 22 continuous process measurements and 19 composition measurements sampled less frequently. The control system selected for the process, a decentralised regulatory control system, was designed by McAvoy and Ye (1994).

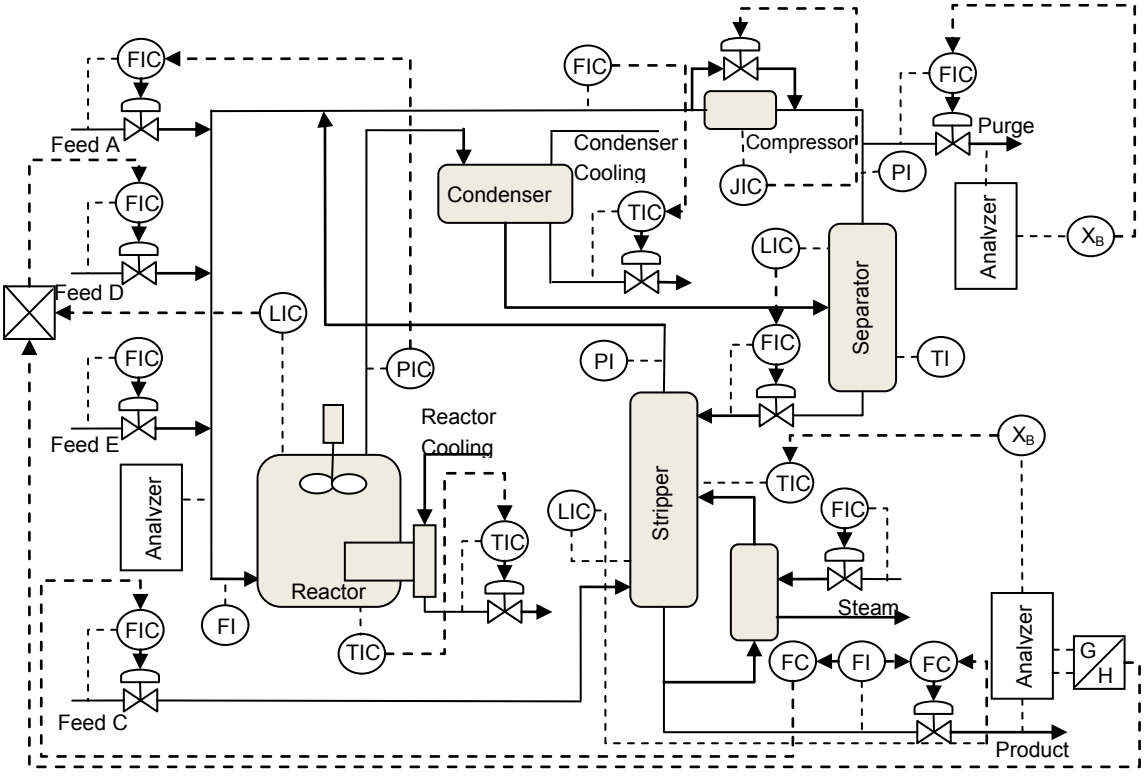

Figure 12: **Control system of the Tennessee Eastman process (adapted from Kano et al., 2002)**

Stellenbosch University http://scholar.sun.ac.za

Listed in Table 1 are all the variables relevant to the Tennessee Eastman process with those normally used for monitoring purposes (Chen and McAvoy, 1998) marked with a (*).

Table 1: **Process variables of the Tennessee Eastman process**

Variable	Description	Variable	Description
1	A_{TE} feed rate*	28	Feed component F_{TE}
2	D_{TE} feed rate*	29	Purge component A_{TE}
3	E_{TE} feed rate*	30	Purge component B_{TE}
4	$A_{TE} + C_{TE}$ feed rate*	31	Purge component C_{TE}
5	Recycle flow*	32	Purge component D_{TE}
6	Reactor feed rate*	33	Purge component E_{TE}
7	Reactor pressure	34	Purge component F_{TE}
8	Reactor level	35	Purge component G_{TE}
9	Reactor temperature*	36	Purge component H_{TE}
10	Purge rate*	37	Product component D_{TE}
11	Separator temperature*	38	Product component E_{TE}
12	Separator level	39	Product component F_{TE}
13	Separator pressure*	40	Product component G_{TE}
14	Separator underflow*	41	Product component H_{TE}
15	Stripper level	42	MV D_{TE} feed flow
16	Stripper pressure*	43	MV E_{TE} feed flow
17	Stripper underflow	44	MV A_{TE} feed flow
18	Stripper temperature*	45	MV $A_{TE} + C_{TE}$ feed flow
19	Stripper steam flow*	46	MV compressor recycle valve
20	Compressor work	47	MV purge valve
21	Reactor cooling water outlet temperature*	48	MV separator underflow
22	Separator cooling water outlet temperature*	49	MV stripper underflow
23	Feed component A_{TE}	50	MV stripper steam valve
24	Feed component B_{TE}	51	MV reactor cooling water flow
25	Feed component C_{TE}	52	MV condenser cooling water flow
26	Feed component D_{TE}	53	MV agitator speed
27	Feed component E_{TE}		

Stellenbosch University http://scholar.sun.ac.za

The set of simulated abnormal conditions consists of (Figure 13):

TEP0 - Normal operating data

TEP1 - A step change in the A_{TE}/C_{TE} feed ratio, B_{TE} composition constant.

TEP2 - A step change in the B_{TE} composition, A_{TE}/C_{TE} ratio constant.

TEP3 - A step change in the D_{TE} feed temperature.

TEP4 - A step change in the reactor cooling water inlet temperature.

TEP5 - A step change in the condenser cooling water inlet temperature.

TEP6 - A step change in the A_{TE} feed loss.

TEP7 - A step change in the C_{TE} header pressure loss - reduced availability.

TEP8 - A random variation in the A_{TE}, B_{TE}, C_{TE} feed composition.

TEP9 - A random variation in the D_{TE} feed temperature.

TEP10 - A random variation in the C_{TE} feed temperature.

TEP11 - A random variation in the reactor cooling water inlet temperature.

TEP12 - A random variation in the condenser cooling water inlet temperature.

TEP13 - A slow drift in the reaction kinetics.

TEP14 - A sticking reactor cooling water valve.

TEP15 - A sticking condenser cooling water valve.

TEP16 - Unknown disturbance.

TEP17 - Unknown disturbance.

TEP18 - Unknown disturbance.

TEP19 - Unknown disturbance.

TEP20 - Unknown disturbance.

TEP21 - A SP change in the production rate (step down 15%).

TEP22 - A SP change in the product mix (50/50 to 40/60).

TEP23 - A SP change in the reactor operating pressure (step down to 60 kPa).

TEP24 - A SP change in the component B_{TE} in purge gas (step up 2%).

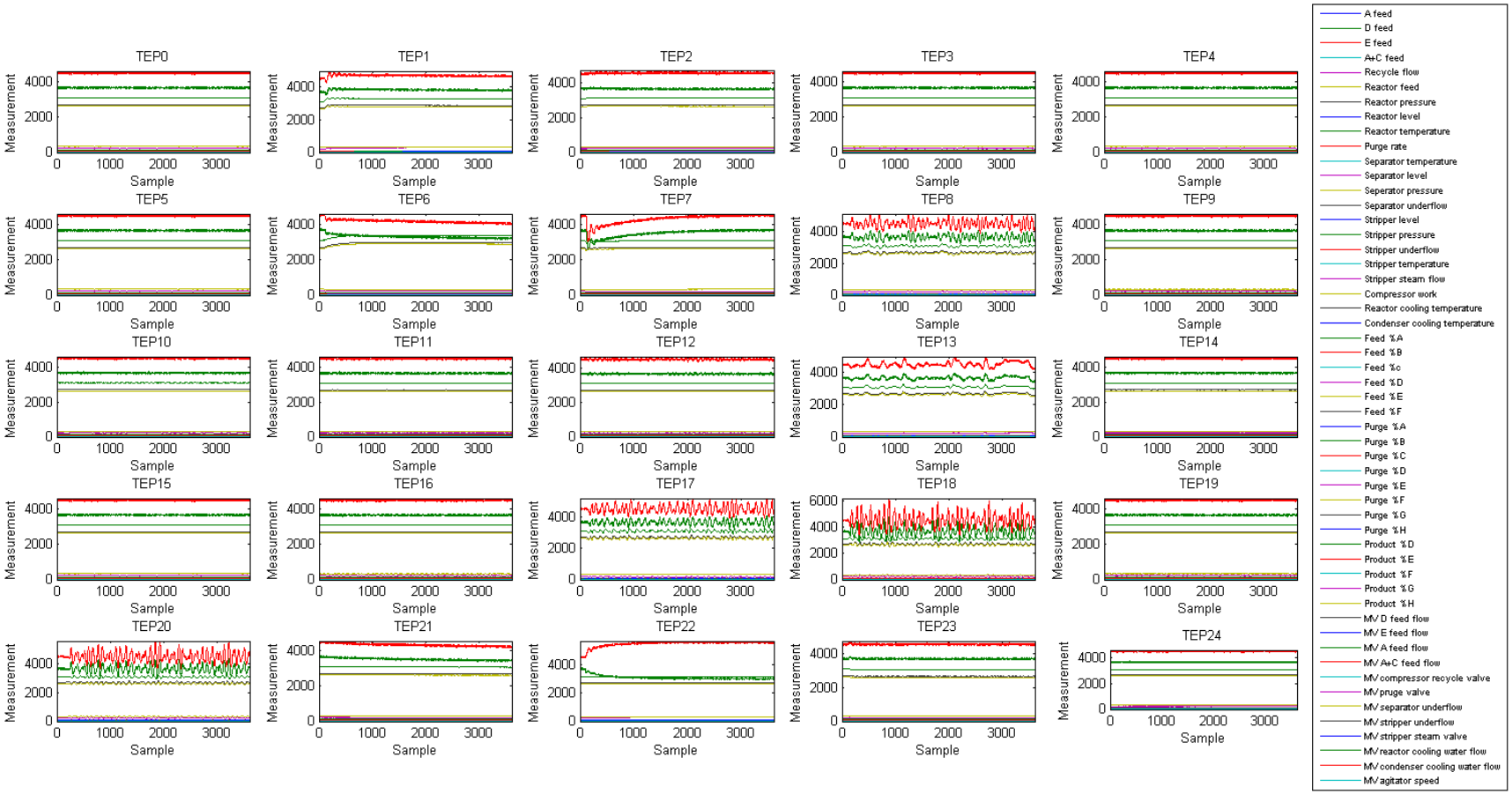

Figure 13: **Tennessee Eastman process: simulated conditions**

Stellenbosch University http://scholar.sun.ac.za

3.6 Technique evaluation

For the evaluation of the statistical data-based fault detection techniques, the 3 data sets / models as described in section 3.5 were used:

- Simple multivariate time series data
- Simple multivariate process (Ku et al., 1995)
- Tennessee Eastman process (Downs and Vogel, 1993)

The evaluation not only focusses on validation of the techniques, but also allows for an appreciation of the techniques and their practicality.

The fault detection performance metrics were tested on both the NOC data as well as all the predetermined fault condition data for each of the data set / models. Listed in Table 2 is an overview of all the performance metrics tested. Since most of the performance metrics being evaluated (Table 2) are PCA-based, all data were scaled to zero mean and unit variance prior to analysis. Furthermore, the variables for each data set were checked to see whether or not the data are independent and identically distributed and have a normal distribution. To test for iid data, each variable was checked to determine if there was any significant autocorrelation in the time series. Significant levels of autocorrelation are indicative of data that is not independent and identically distributed. To test if the data is normally distributed, each variable was subjected to the Lilliefors test. The Lilliefors test is a goodness-of-fit test of composite normality, testing whether or not the data in a time series come from an unspecified normal distribution. Such tests determining the statistical characteristics of the data is critical in that it allows appropriate techniques, having specific assumptions, to be matched up with the data.

Each monitoring method can be evaluated using the following procedure (Kano et al., 2002):

- Each monitoring method is applied to the data of both the normal operating condition and all of the abnormal conditions, and the monitoring indices calculated.
- After the occurrence of an abnormal condition, the percentage of samples outside the control limit is calculated, termed *reliability*. This is typically done for the first 100 samples following the occurrence of an abnormal condition.
- The average reliability of a number of simulations having different initial conditions is subsequently calculated in each case.

For control limits representing 99% confidence limits, a monitoring method having a reliability considerably higher than 90% can be regarded as successful in detecting abnormal conditions. However, for reliabilities less than or close to 90%, the monitoring method can be regarded as not functioning well. It should be noted that the reliability is affected by the number of samples used for calculating it.

Other, useful performance measures also calculated include:

- The percentage of samples outside the control limit for a known NOC data set, termed *false alarm rate*. De Gooijer (2006) has found that all test statistics generally become more effective as the data dimension increases, in turn ensuring that not too many false alarms are raised.

Stellenbosch University http://scholar.sun.ac.za

- The number of faulty samples, following the occurrence of an abnormal condition, before the abnormal condition is detected, termed *detection-delay*.

Table 2: **Overview of statistical data-based fault detection techniques**

Technique	Type	Metric	Comments
Shewhart	Univariate	Process data	Effective at detecting large changes in the mean of data.
CUSUM	Univariate	CUSUM of the deviation of the process data	Effective at detecting small changes in the mean of data.
EWMA	Univariate	EWMA of the process data	Adds time history of data to detect changes in the mean of data.
PCA	Multivariate	SPE T^2	Linear technique. Assume data to be normally distributed.
Moving PCA	Multivariate	A_i A_{1-m}	Focus on changes in the distribution of the data. Better at detecting small changes compared to PCA.
DISSIM	Multivariate	Dissimilarity index	Focus on changes in the distribution of the data. Sensitive to changes in the data correlation structure.
DPCA	Dynamic multivariate	SPE T^2	Exploit autocorrelation in data through use of time lag shift method. Better at detecting small changes compared to PCA.
SSPCA	Dynamic multivariate	SPE T^2	Increase in resolution over PCA, although having a delayed reaction. Better at detecting small changes compared to PCA.
MSPCA	Dynamic multivariate	SPE T^2	Model data containing contributions from events whose behaviour changes over time and frequency
NNPCA	Non-linear multivariate	SPE T^2	Non-linear technique. No assumption regarding data distribution.
AANNPCA	Non-linear multivariate	SPE	Improved separation of non-linear factors in the data compared to NNPCA. No assumption regarding data distribution.
NLPCA	Non-linear multivariate	SPE	Not robust against outliers.
ELM PCA	Non-linear multivariate	SPE T^2	Extremely fast learning speed and significantly less parameters needing estimation compared to NNPCA. Collinearity potentially a problem.

Stellenbosch University http://scholar.sun.ac.za

3.6.1 Simple multivariate time series data

For the simple multivariate time series data 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. The following is a list of analysis configuration parameters used for the various performance metrics:

- Confidence limit threshold = 99%
- CUSUM windows size = 60 seconds
- Moving PCA window size = 60 seconds
- Dissimilarity index window size = 60 seconds
- Summed-scores window size = 60 seconds
- Multiscale PCA wavelet level of decomposition = 5
- PCA number of principal components selected = 3 (explaining 87.6% of the variance)
- DPCA embedding dimension = 3
- DPCA number of principal components selected = 9 (explaining 88.2% of the variance)

Inspection of the simple multivariate time series data (Figure 10) indicated it to be independent and identically distributed while having a normal distribution. These data characteristics meet the required assumptions of all of the fault detection performance metrics being evaluated; ensuring their relevance to the data set being analysed. Furthermore, since the data set comprised entirely of randomly generated data with only a predefined covariance between x_2, x_3 and x_4 it is expected that, although applicable, there would be no need for the additional complexity offered by the non-linear and dynamic multivariate performance metrics. For the sake of completeness, however, all the available metrics will be assessed. This will not only allow all appropriate performance measures to be evaluated but also help to gain insight into the potential benefits that may be offered by the more complex performance measures in terms of reduced false alarm rates, improved reliability, or reduced detection-delays.

From the false alarm rate evaluation graph (Figure 14) it can be seen that most performance metrics have the expected false alarm rate of approximately 1, given the confidence limit threshold of 99%. The only exception to this is the performance metrics using neural networks (including ELM) for modelling the data, where the increase in the false alarm rate can be ascribed to the modelling error introduced by the neural networks. Whereas for normal PCA the input data is exactly modelled, for neural network PCA, a neural network is first used to generate residuals from the input data, introducing a modelling error, prior to PCA being applied. For the auto-associative neural network PCA, the first non-linear score is derived from the main data with subsequent non-linear scores being derived from the model residuals, again introducing a modelling error. This results in the input data not being exactly modelled for auto-associative neural network PCA. It is interesting to note that for the performance metrics using neural networks for modelling the data, the model error when applying the model to new data, also gives an indication as to

whether or not a fault condition exists in the data with a low R^2 being indicative of a fault condition and a high R^2 indicating NOC data.

For the reliability index shown in Figure 15, the maximum value obtained for each of the performance metric and fault condition combinations evaluated is indicated by the colour of the square. The detection-delay corresponding to the most reliable performance metric is subsequently shown in Figure 16. A performance metric is considered to be performing well if it has a reliability index of greater than 90%.

For the step change fault conditions, MTS1-MTS4, most of the techniques were able to easily identify a fault condition in the data set. All the univariate statistical process monitoring performance metrics, Shewhart, CUSUM and EWMA were always able to correctly identify the faulty variable, with the auto-associative neural network PCA, principal curves PCA and summed-scores individual variable SPE being able to correctly identify the faulty variable for small changes, but less so for large changes. Some of the performance metrics were only able to identify a fault condition in the data set once the magnitude of the fault condition exceeded some threshold.

For the ramp change fault conditions, MTS5-MTS8, all the techniques were slow to detect a fault condition in the data as is evident from the detection-delay results (Figure 16). The primary reason for this is the fact that the ramp change fault condition was taking quite a bit of time to exceed the performance metrics detection thresholds, however most techniques showed great potential in eventually identifying a fault condition in the data set given enough time. Again, all the univariate statistical process monitoring performance metrics, Shewhart, CUSUM and EWMA were always able to correctly identify the faulty variable, although never exceeding 90% for the reliability index, with the principal curves PCA, dynamic PCA and summed-scores individual variable SPE being able to correctly identify the faulty variable for small changes, but not for large changes, again never exceeding 90% for the reliability index.

For the spike fault conditions, MTS9-MTS12, all the techniques performed poorly when looking at the reliability index. The primary reason for this is the fact that the reliability index is measured over a time span of 100 seconds whereas the spike fault condition only last between 1 and 10 seconds. However, the short term results immediately following the fault condition are similar to those obtained for the step change fault condition. The high detection-delay values indicated in Figure 16 for these fault conditions are all due to false alarms being detected related to performance metrics correctly having very low reliability index values.

For the standard deviation change fault conditions, MTS13-MTS16, moving PCA A_i, moving PCA A_{1-m} and the dissimilarity index performance metrics performed exceptionally well in identify a fault condition in the data for the uncorrelated variable, with a few other performance metrics also showing potential for detecting larger fault condition changes. For a change in the correlated variable, moving PCA A_{1-m} and the dissimilarity index performance metrics performed exceptionally well in identifying a fault condition in

Stellenbosch University http://scholar.sun.ac.za

the data, again with a few other performance metrics showing potential for detecting larger fault condition changes.

For the covariance shift fault conditions, MTS17-MTS19, in addition to the dynamic PCA SPE and dynamic PCA T^2 performance metrics being able to detect a fault condition in the data set, moving PCA A_{1-m} and the dissimilarity index performance metrics again performed exceptionally well. A few of the other performance metrics also showed potential for detecting larger covariance changes in the data set.

For the function change fault conditions, MTS20-MTS23, as with the covariance shift fault conditions, moving PCA A_{1-m}, the dissimilarity index, dynamic PCA SPE and the dynamic PCA T^2 performance metrics all effectively detected a fault condition in the data set. The PCA T^2, neural network PCA T^2, extreme learning machine PCA T^2 and multiscale PCA T^2 performance metrics was also effective in detecting a change in the data set for both a change in the uncorrelated and correlated variable with many of the other performance metrics also being able to detect larger changes in the correlated variable.

From this evaluation (Figure 15) it can be concluded that the univariate techniques were the most reliable to detect “simple” changes in the data (step change, ramp change and spike) with the basic multivariate techniques being sufficient to detect the more “complex“ changes in the data (standard deviation change, covariance shift and function change). Furthermore, it can be concluded that for the simple multivariate time series data, being independent and identically distributed with a normal distribution, as expected the basic multivariate performance metrics, namely the moving PCA A_{1-m} and dissimilarity index performance measures, outperformed all the other performance measures on average for all the fault conditions tested. On average, the best extreme learning machine PCA performance measure had an average reliability of 71.56% over all fault conditions and was found to be within with the top 20% of performance measures evaluated, easily outperforming the neural network PCA performance measures. For this case study there was little, if any, benefit to applying more complex fault detection techniques to the data.

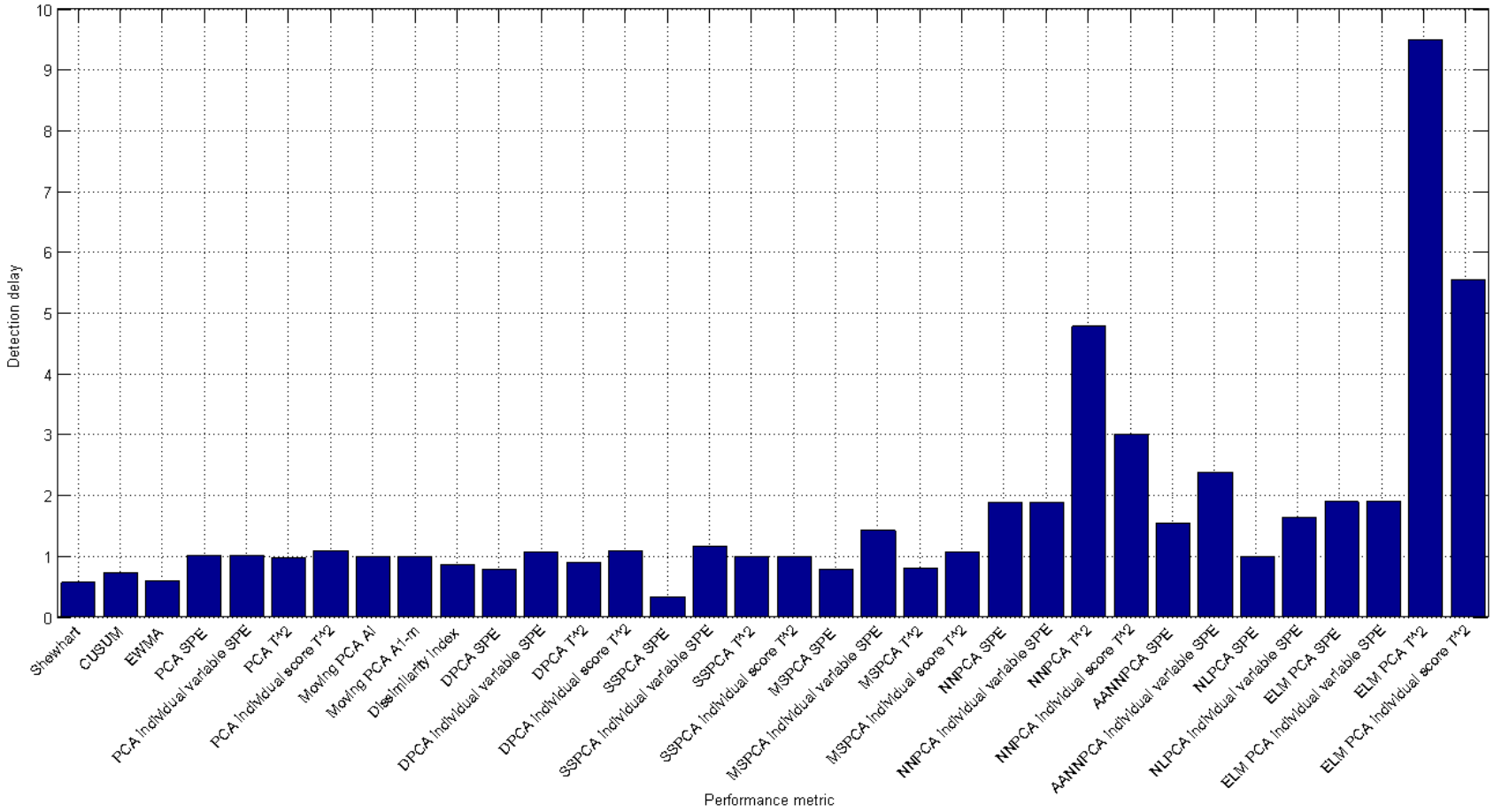

Figure 14: **Simple multivariate time series data: fault detection false alarm rates at a confidence level of 0.99**

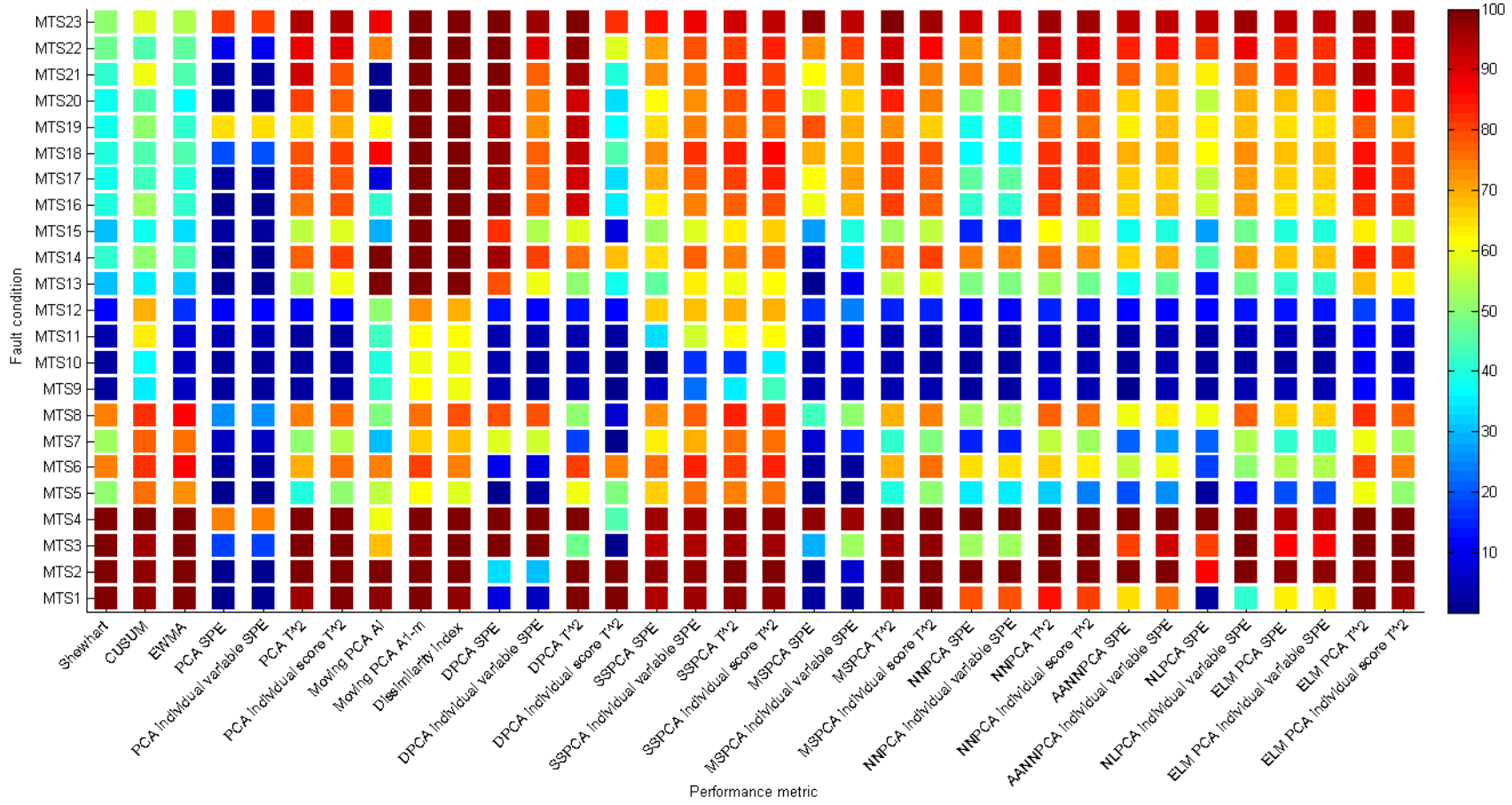

Figure 15: **Simple multivariate time series data: fault detection reliability index at a confidence level of 0.99**

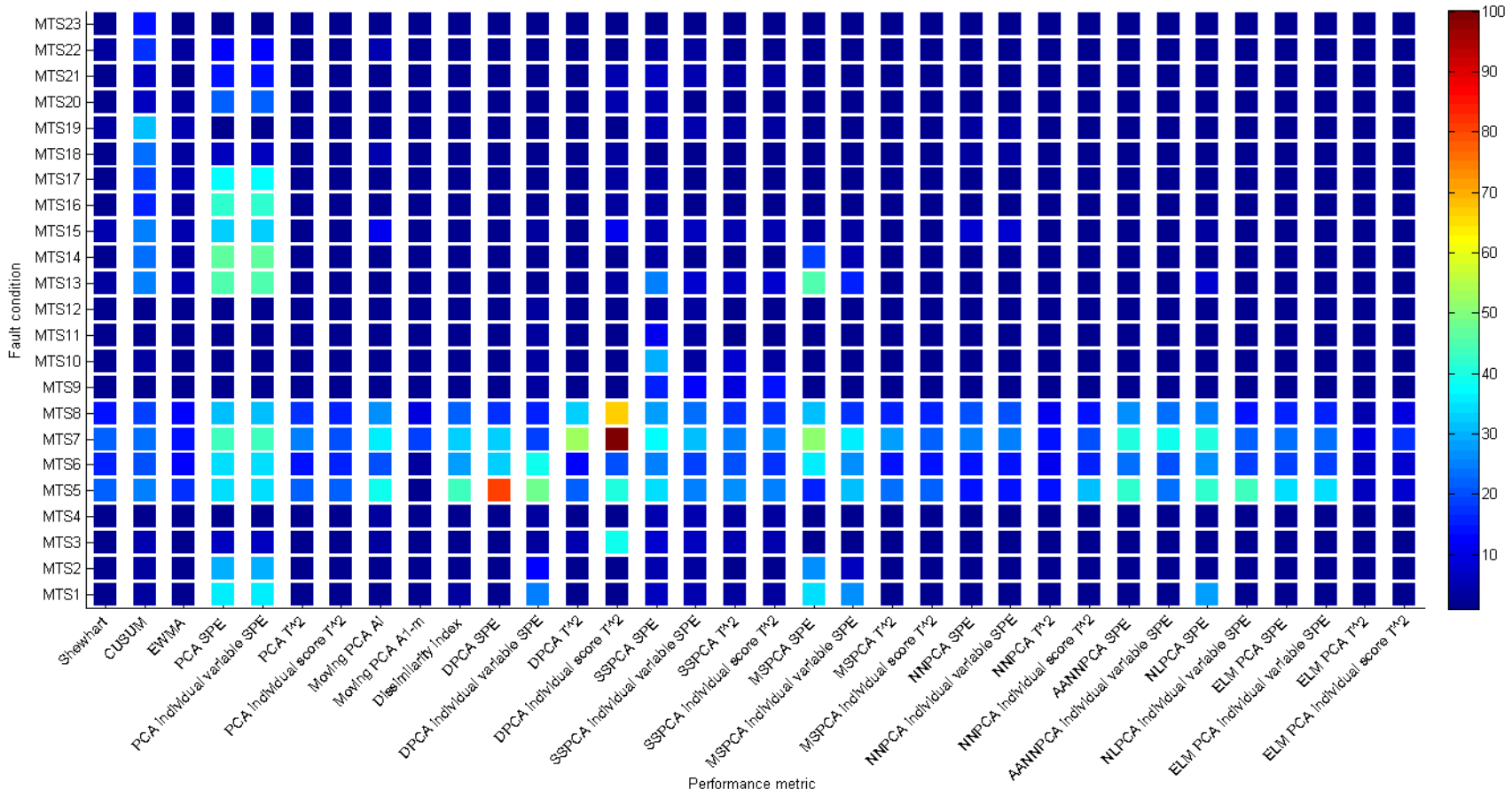

Figure 16: **Simple multivariate time series data: fault detection detection-delay at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

3.6.2 Simple multivariate process

For the simple multivariate process 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. The following is a list of analysis configuration parameters used for the various performance metrics:

- Confidence limit threshold = 99%
- CUSUM windows size = 60 seconds
- Moving PCA window size = 60 seconds
- Dissimilarity index window size = 60 seconds
- Summed-scores window size = 60 seconds
- Multiscale PCA wavelet level of decomposition = 5
- PCA number of principal components selected = 2 (explaining 82.6% of the variance)
- DPCA embedding dimension = 3
- DPCA number of principal components selected = 3 (explaining 81.6% of the variance)

Inspection of the simple multivariate process data (Figure 11) indicated it to not be independent and identically distributed while having a normal distribution. These data characteristics violate the required assumptions of the univariate and most of the non-linear multivariate performance metrics, making them unsuitable for assessing this particular data set. However, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Since the data for this case study is based on a simple multivariate process containing correlated inputs and including history values, it is expected that the non-linear and dynamic multivariate performance metrics may be required for reliable fault detection. However, the magnitudes of the process changes resulting in the fault conditions are very small and it is expected that all the available techniques will struggle to reliably detect the fault conditions. For the sake of completeness all the available metrics will therefore be assessed. This will not only allow all appropriate performance measures to be evaluated but also help to gain insight into the robustness of the performance metrics whose data assumptions have not been met.

From the false alarm rate evaluation graph (Figure 17) it can be seen that most performance metrics have the expected false alarm rate of approximately 1, given the confidence limit threshold of 99%. As with the simple multivariate time series case study (3.6.1), the performance metrics using neural networks (including ELM) for modelling the data again showed a slightly larger false alarm rate. In addition to this, the EWMA performance metrics also showed a slightly larger false alarm rate. This is directly related to the exponential filter coefficient, determining the memory length, used by the EWMA performance metric. Whereas increasing the exponential filter coefficient reduces the false alarm rate, decreasing the exponential filter coefficient increases the false alarm rate, both changes affecting the reliability of the performance metric.

Stellenbosch University http://scholar.sun.ac.za

For the reliability index shown in Figure 18, the maximum value obtained for each of the performance metric and fault condition combinations evaluated is indicated by the colour of the square. The detection-delay corresponding to the most reliable performance metric is subsequently shown in Figure 19. A performance metric is considered to be performing well if it has a reliability index of greater than 90%.

For the mean shift fault conditions, SMP1-SMP5, only a select few of the performance metrics were able to detect a fault in the data set, and only for a significant mean shift, never exceeding 90% for the reliability index. As the magnitude of the mean shift increased, the CUSUM, EWMA and summed-scores individual variable SPE performance metrics started being able to detect the fault condition in both the output and input variables more reliable with the summed-scores SPE, summed-scores T^2 and summed-scores individual score T^2 being able to detect a general fault condition in the data, none of which ever exceeded 90% for the reliability index. The reason for the success of the summed-scores PCA performance metrics can be ascribed to the fact that the time-dependent relationships among the variables are accounted for using this technique. In general it was also found that as the magnitude of the mean shift increased, the reliability index of the performance metrics increased and their associated detection-delays decreased (Figure 19).

For the parameter change fault conditions, SMP6-SMP8, only the moving PCA A_{1-m} performance metric was really able to detect a fault condition in the data set, although never exceeding 90% for the reliability index. The reason for the success of the moving PCA A_{1-m} performance metric in detecting the fault condition can be ascribed to the fact that this metric is particularly suited to detecting changes in variables characterised by changes in the correlation structure among variables. Since the variances of the principal components are similar for moving PCA in this case, the moving PCA A_i performance metric is ineffective. This inadequacy is, however, addressed by the moving PCA A_{1-m} performance metric. As with the mean shift fault condition it was found that as the magnitude of the change of parameters increased, the reliability index of the performance metrics increased.

As expected, due to the small magnitudes of the process changes resulting in the fault conditions, most the available techniques struggled to reliably detect the fault conditions. From this evaluation (Figure 18), although the univariate techniques (CUSUM and to a lesser degree the EWMA) were sufficient to detect “simple” changes (mean shift) in the data, the dynamic multivariate summed-scores performance metrics were slightly more reliable, with the basic multivariate moving PCA A_{1-m} being the only technique able to detect the more “complex“ changes (change of parameters) in the data. It should be noted that it has been mentioned in the literature that these three techniques are all effective at detecting small shifts in data when compared to the other techniques being evaluated. It can therefore be concluded that for the simple multivariate process data, not being independent and identically distributed with a normal distribution, the multivariate moving PCA A_{1-m} performance metric is the most reliable for detecting step

Stellenbosch University http://scholar.sun.ac.za

changes with the dynamic multivariate summed-scores performance metrics being the most reliable for detecting parameter changes in the data. On average, the best extreme learning machine PCA performance measure had an average reliability of only 16.90% over all fault conditions but was found to still be within with the top 25% of performance measures evaluated, again outperforming the neural network PCA performance measures. For this case study, as with the previous one, it was found that there was little, if any, benefit to applying more complex fault detection techniques to the data.

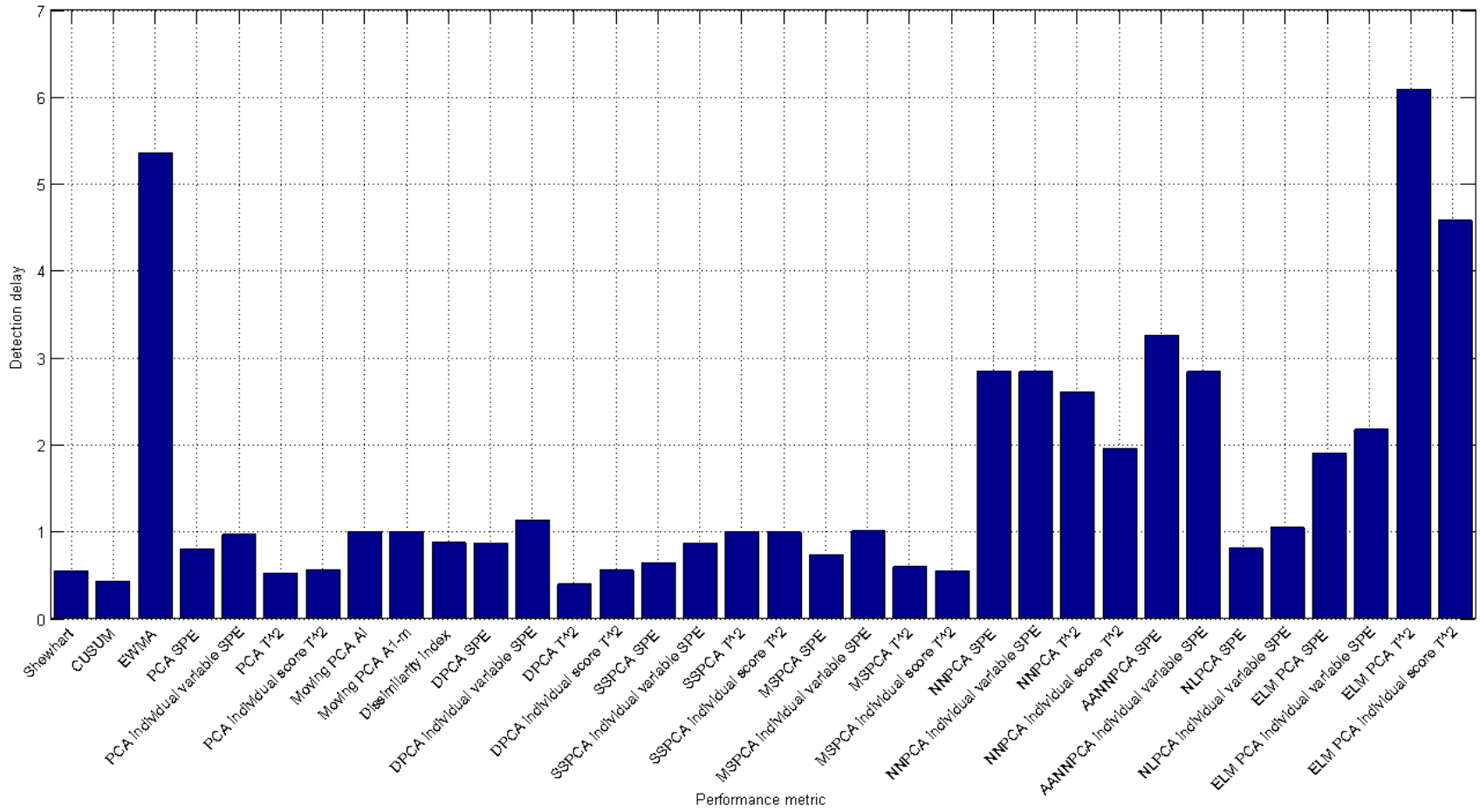

Figure 17: **Simple multivariate process: fault detection false alarm rates at a confidence level of 0.99**

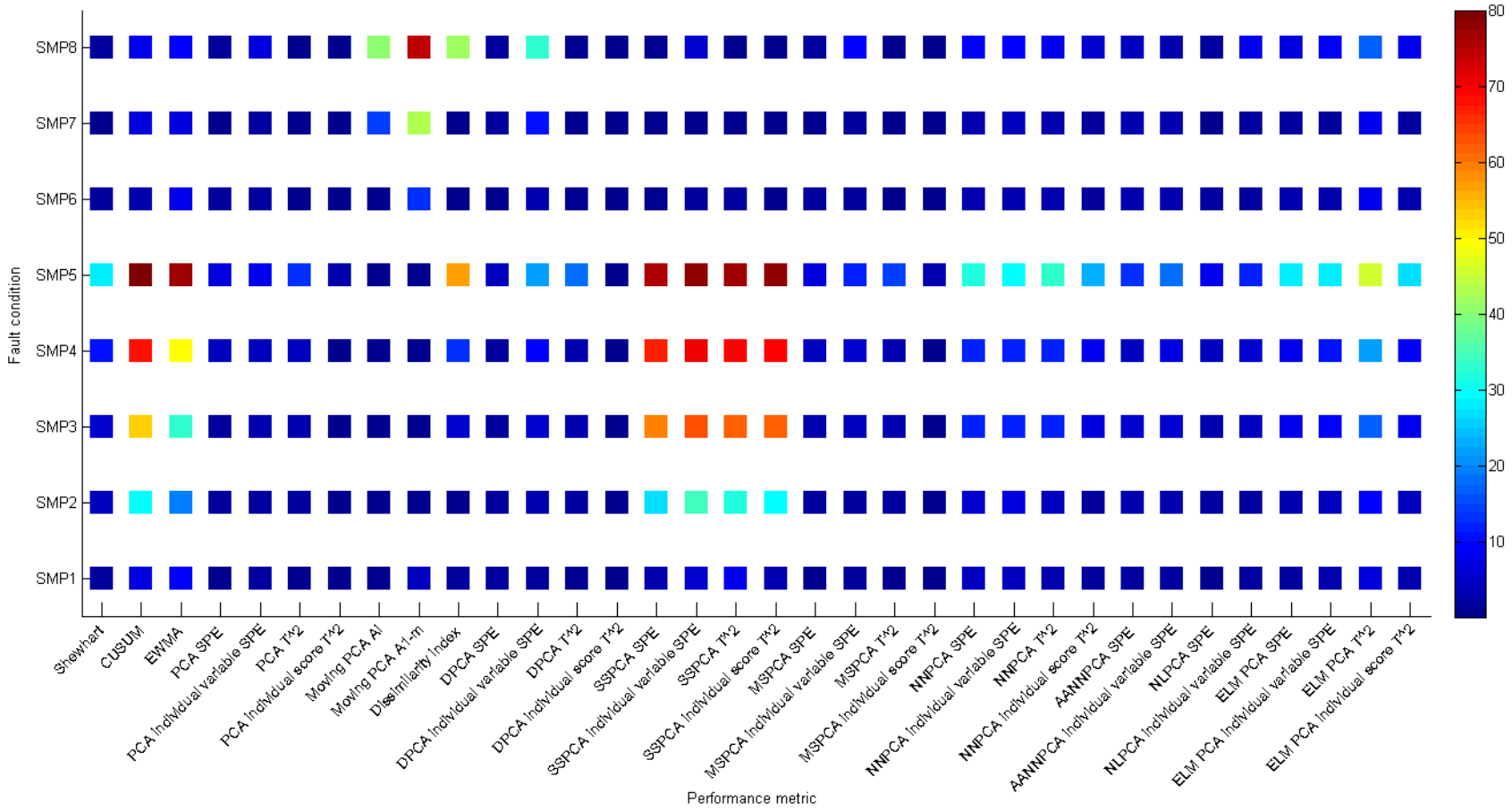

Figure 18: **Simple multivariate process: fault detection reliability index at a confidence level of 0.99**

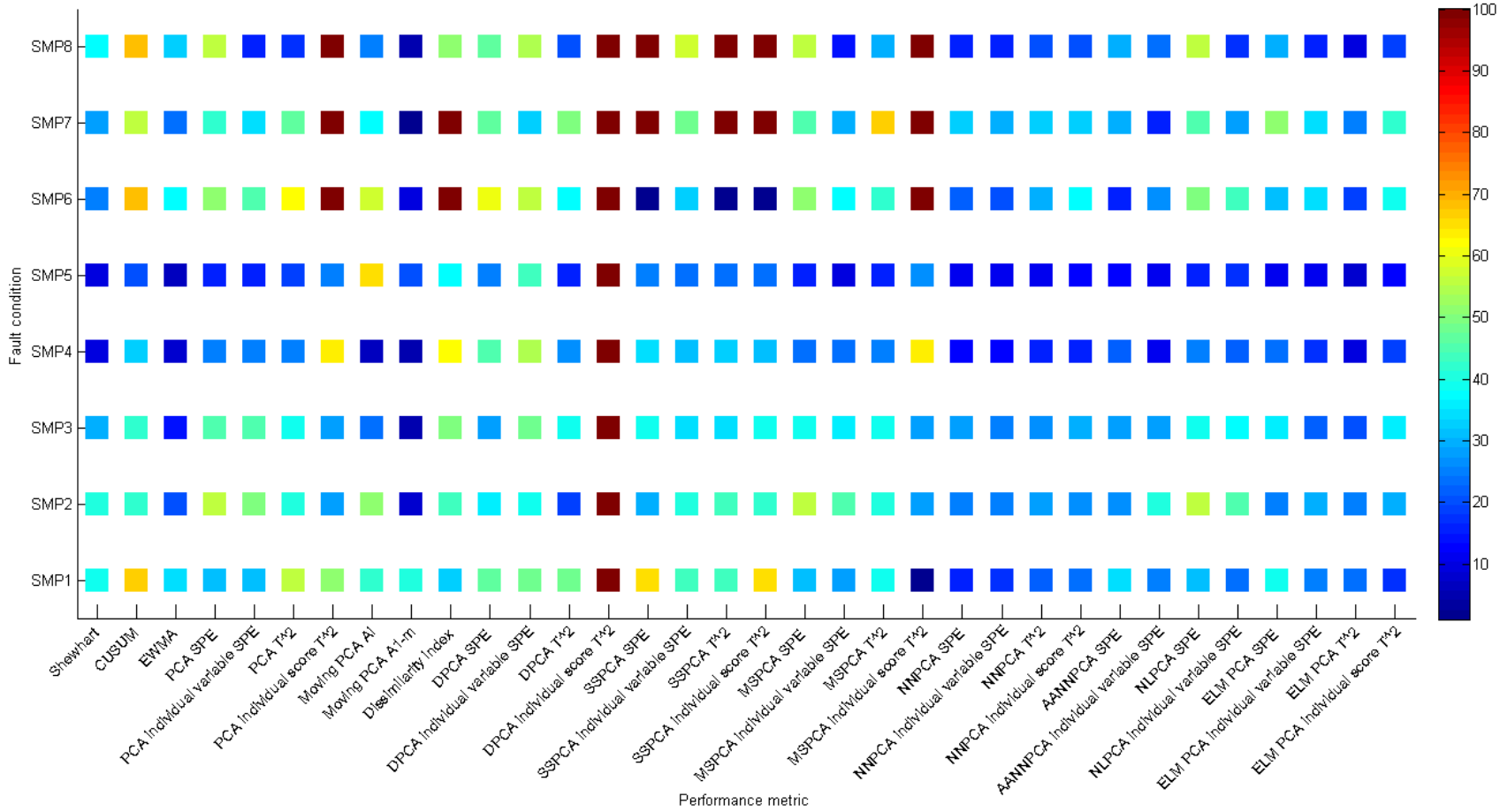

Figure 19: **Simple multivariate process: fault detection detection-delay at a confidence level of 0.99**

3.6.3 Tennessee Eastman process

For the Tennessee Eastman process 3600 data points at a sampling rate of one sample per ninety seconds were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. The following is a list of analysis configuration parameters used for the various performance metrics:

- Confidence limit threshold = 99%
- CUSUM windows size = 5 minutes
- Moving PCA window size = 5 minutes
- Dissimilarity index window size = 5 minutes
- Summed-scores window size = 5 minutes
- Multiscale PCA wavelet level of decomposition = 5
- PCA number of principal components selected = 9 (explaining 81.8% of the variance)
- DPCA embedding dimension = 4
- DPCA number of principal components selected = 25 (explaining 80.1% of the variance)

Inspection of the Tennessee Eastman process data (Figure 13) indicated it to have a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, with many of the variables having significant autocorrelation with similar profiles. This similarity in the autocorrelation profiles is indicative of the presence of high levels of collinearity (linear correlation structure between the variables). Based on these data characteristics, none of the performance metrics would be suitable for assessing all of the data in this particular data set. However, as stated earlier, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Furthermore, with the data for this case study being based on a complex multivariate process it is expected that the more complex performance metrics, such as the non-linear multivariate performance metrics, be required for effective fault detection. Considering that this is quite an extensive data set it is expected that some of the fault conditions will relate to data that is independent and identically distributed while having a normal distribution, resulting in some of the less complex performance metrics being valid some of the time. Therefore, all the available metrics will be assessed. This will not only allow all appropriate performance measures to be evaluated but also help to gain insight into the robustness of the performance metrics whose data assumptions have not been met.

From the false alarm rate evaluation graph (Figure 20) it can be seen that most performance metrics have the expected false alarm rate of approximately 1, given the confidence limit threshold of 99%. As with the simple multivariate process case study (3.6.2), both the EWMA performance metrics and the performance metrics using neural networks (including ELM) for modelling the data again showed a slightly larger false alarm rate.

Stellenbosch University http://scholar.sun.ac.za

For the reliability index shown in Figure 21, the maximum value obtained for each of the performance metric and fault condition combinations evaluated is indicated by the colour of the square. The detection-delay corresponding to the most reliable performance metric is subsequently shown in Figure 22. A performance metric is considered to be performing well if it has a reliability index of greater than 90%.

From the reliability index results (Figure 21) the simulated fault conditions can be divided into 3 distinct groups:

a) Fault conditions that were easily detected by the majority of the performance metrics: fault conditions TEP1, TEP6, TEP7, TEP13, TEP17, TEP22, TEP23 and TEP24. Detection-delays for this group of fault conditions where typically very low (Figure 22).
b) Fault conditions that were typically detected by approximately half of the performance metrics: fault conditions TEP2, TEP8, TEP10, TEP11, TEP16, TEP18 and TEP20. Detection-delays for this group of fault conditions where typically very low (Figure 22).
c) Fault conditions that were not reliably (reliability index of greater than 90%) detected by the majority of the performance metrics: fault conditions TEP3, TEP4, TEP5, TEP9, TEP12, TEP14, TEP15, TEP19 and TEP21. Detection-delays for this group of fault conditions where typically very high and more often than not related to false alarms (Figure 22).

The fault conditions in group (a) consist mainly of step changes, with a single ramp parameter change and an “other” disturbance. For this group typically either the variable responsible for the fault condition or a variable closely related to the fault condition is being monitored. Also taking into account the fact that the majority of the fault conditions in this group can be seen as “basic”, most of the performance metrics were very reliable in detecting these fault conditions except for the CUSUM, PCA individual score T^2, moving PCA A_i, moving PCA A_{1-m}, dissimilarity index and dynamic PCA individual score T^2 performance metrics. It was further found that for fault conditions detected by the Shewhart or EWMA performance metrics, the metrics were also very effective in identifying the relevant affected variables.

The fault conditions in group (b) consist mainly of random variation changes and “other” disturbances, with a single step change fault condition. Unlike for group (a) fault conditions, for group (b) the variable responsible for the fault condition was rarely directly being monitored, however, variables related to the fault conditions were mostly being monitored. Also taking into account the fact that the majority of the fault conditions in this group can be seen as “advanced”, univariate and basic multivariate performance metrics, except PCA SPE and PCA individual variable SPE, were very unsuccessful in detecting the fault conditions. However, all non-linear multivariate performance metrics, except extreme learning machine PCA individual score T^2, and most dynamic multivariate performance metrics, except dynamic PCA T^2, dynamic PCA individual score T^2 and all summed-scores performance metrics, were very reliable at detecting the fault conditions. It was further found that for fault conditions detected by the principal curves PCA and multiscale PCA individual variable SPE performance metrics, the metrics were also very effective in identifying the relevant affected variables.

Stellenbosch University http://scholar.sun.ac.za

The fault conditions in group (c) consist of a mixed bag of step changes, random variation changes and "other" disturbances. For this group, none of the performance metrics were able to reliably detect any of the fault conditions. This can be ascribed to the fact that variables responsible for the fault conditions or variables closely related to the fault conditions were not being monitored (Chen and McAvoy, 1998), the control system controlling the Tennessee Eastman process was too quick in correcting the induced fault condition, the effect of the fault condition on the process was too slow or the magnitude of the fault condition was too small to result in a measurable process performance degradation over the evaluation period (variation still within variation associated with common cause).

From this evaluation (Figure 21) it can be concluded that the principal curves PCA (non-linear multivariate) and multiscale PCA (dynamic multivariate) individual variable SPE performance metrics where not only very successful in detecting the various fault conditions but also the most accurate in identifying possible affected variables. It should also be noted that the neural network PCA performance metrics were the most reliable in detecting the fault conditions listed for group (a) and (b). It can therefore be concluded that for the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, the non-linear multivariate neural network PCA performance measures outperformed all the other performance measures on average for all the fault conditions tested. Although the data characteristics violated the assumptions of all the fault detection techniques tested, by far the majority of the techniques proved to be very robust in this regard and were still able to detect many of the fault conditions. On average, the best extreme learning machine PCA performance measure had an average reliability of 71.42% over all fault conditions and was found to be within with the top 10% of performance measures evaluated, only marginally being outperformed by two of the neural network PCA performance measures. For this case study it was therefore found that more complex fault detection techniques were critical to the detection of fault condition in the Tennessee Eastman process. One approach to potentially reducing the complexity inherent to monitoring the Tennessee Eastman process as a whole lies with the idea of simplifying the problem through the use of process causality maps.

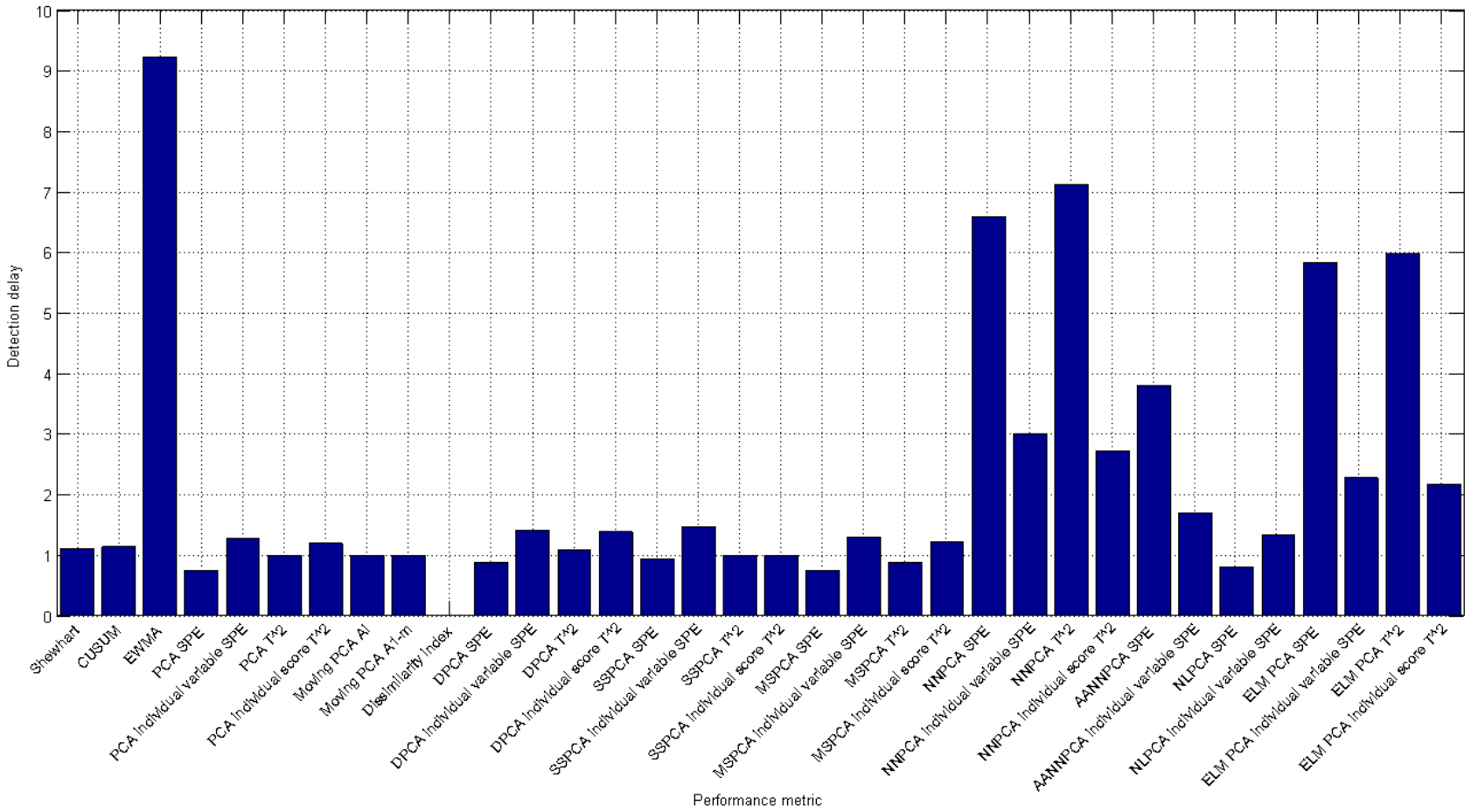

Figure 20: **Tennessee Eastman process: fault detection false alarm rates at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

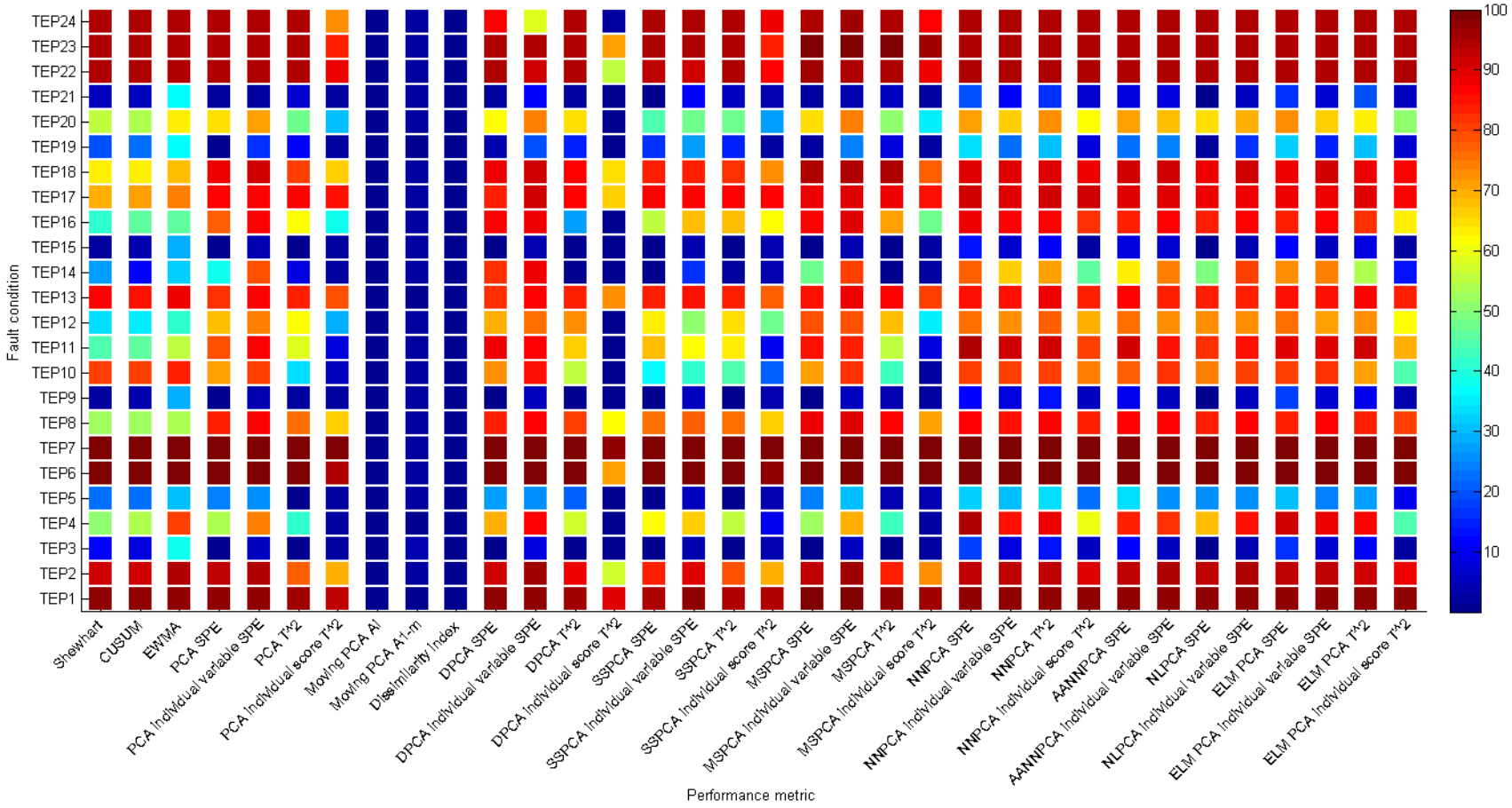

Figure 21: **Tennessee Eastman process: fault detection reliability index at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

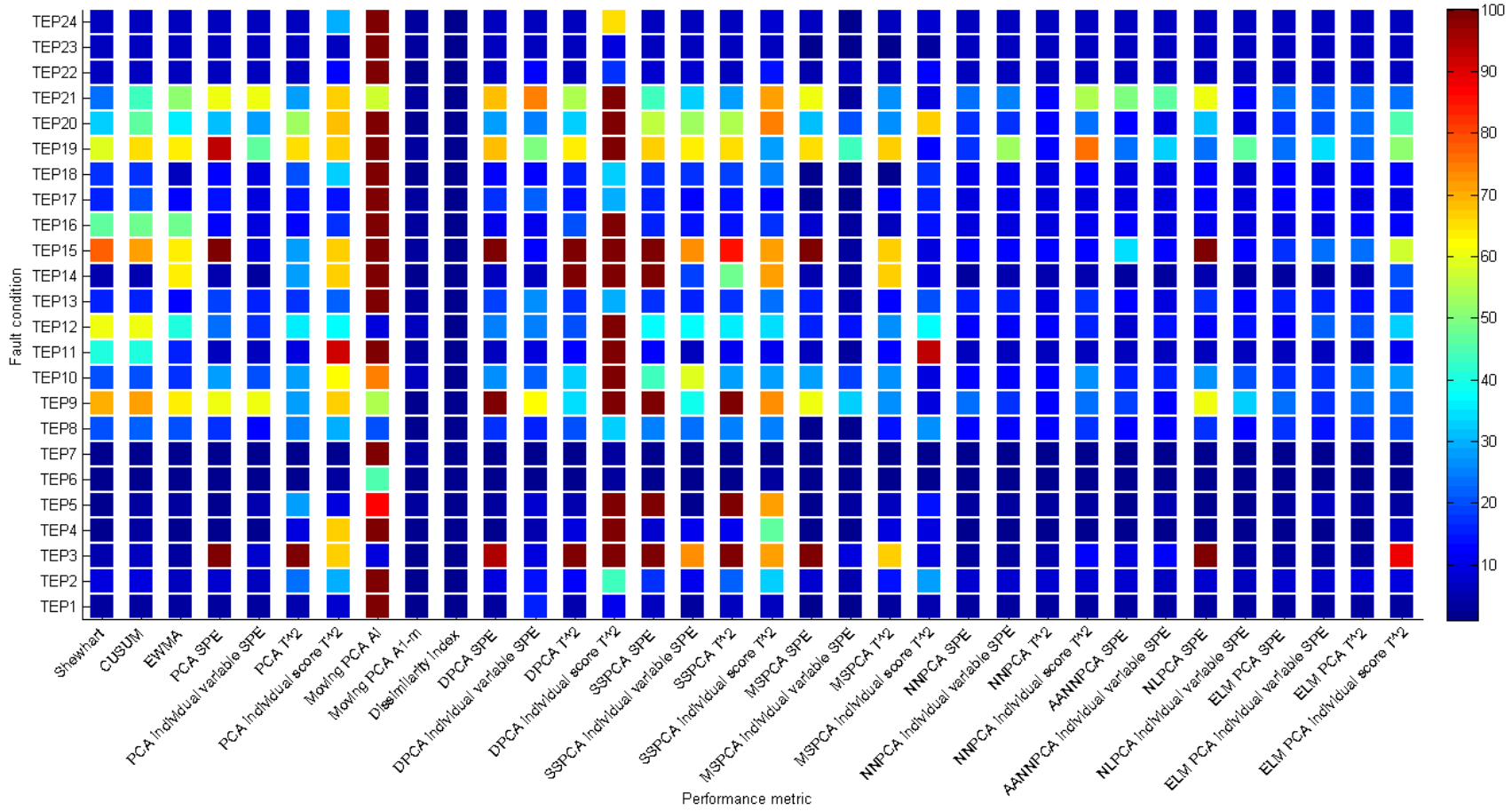

Figure 22: **Tennessee Eastman process: fault detection detection-delay at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

3.6.4 Summary

From the fault detection technique evaluation it is evident that there is no single fault detection technique that is effective in detecting all potential fault conditions. Whereas univariate and basic multivariate performance metrics where very effective in detecting fault conditions in all the case studies, many fault conditions for the Tennessee Eastman process case study required non-linear multivariate or dynamic multivariate performance metrics to detect reliably. This confirms the fact that the different fault detection techniques are suited better to different types of data structures and specialised to detect very specific fault conditions:

- For the simple multivariate time series data, being independent and identically distributed with a normal distribution, the basic multivariate performance metrics, namely the moving PCA A_{1-m} and dissimilarity index performance measures, were found to be the most reliable.
- For the simple multivariate process data, not being independent and identically distributed with a normal distribution, the basic multivariate moving PCA A_{1-m} and dynamic multivariate summed-scores performance measures were found to be the most reliable.
- It was further confirmed that the CUSUM, moving PCA and summed-scores performance measures were all more effective at detecting small shifts in data when compared to the other techniques being evaluated.
- For the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, the non-linear multivariate neural network PCA performance measures were found to be the most reliable.
- It was further found that even when the data characteristics violate the assumptions underpinning a fault detection technique, often the technique will prove to be very robust in this regard and still be able to detect many of the fault conditions presented to it.

For all the performance metrics evaluated, except the EWMA and those utilising neural networks, a very low false alarm rate was obtained while still achieving an acceptable reliability. Combined with the fact that no single fault detection technique was found to be effective in detecting all potential fault conditions it is suggested that multiple fault detection techniques be run in parallel and if any one of them detects a fault condition the event be investigated. This will allow the benefits that each of the different performance metrics offers to be leveraged simultaneously and in so doing ensuring that the combined performance metrics outperform any of the individual performance metrics when presented with a variety of different fault conditions.

Following the success of the non-linear multivariate neural network PCA performance measures (Chen and Liao, 2002), it was proposed to replace the neural network algorithm with the ELM algorithm in order to remove the non-linear and dynamic characteristics from the data prior to the application of PCA. This decision was based on the advantages the ELM algorithm has over conventional feedforward neural network algorithms: being good at generalisation, having extremely fast learning speeds and having very few parameters that require setting (Huang et al., 2006). For the proposed ELM PCA algorithm it was

Stellenbosch University http://scholar.sun.ac.za

found that its best performing performance measure was consistently in the top 25% of performance measures evaluated. Relative to the other performance measures, the algorithm also performed exceptionally well for the Tennessee Eastman process case study, which can be ascribed to the non-linear multivariate nature of the underlying model. It was further found that the ELM PCA derived performance measures typically outperformed the NNPCA derived performance measures, attributed to its good generalization performance, except for the Tennessee Eastman process case study where their performance were very similar, possibly due to overfitting during training.

For complex processes such as the Tennessee Eastman process it is further suggested to reduce the complexity of the process through the development of process causality maps. Process causality maps should simplify the challenge of monitoring the process by reducing it to manageable portions. This will allow multiple, smaller, individual processes to be monitored at a low level, while still allowing the overall process to be monitored at a higher level. Not only should this improve the reliability of the fault detection results through a more focussed application, but the interpretability of the results should also improve due to the reduction in complexity.

Stellenbosch University http://scholar.sun.ac.za

4 Change point detection

Statistical data-based fault detection techniques are typically specifically applied for monitoring process deviations, focussing mainly on detecting abnormal process conditions compared to normal operating process data. On the other hand there are instances when wanted or unwanted change points can occur while the process is in a state of either normal or abnormal process operation. Change point detection complements statistical data-based fault detection techniques in that it aims to determine whether or not a process in its current state is exhibiting a change in behaviour compared to the behaviour it exhibited immediately preceding its current state, irrespective of whether it was in a state of normal process behaviour or not.

Change points in data can be viewed as times of discontinuities induced for example from changes in a process, input conditions, equipment and/or measurement techniques. Locating change points can be seen as the equivalent of finding the optimal way to partition time series data, splitting the data in different ways and measuring the divergence between the partitions by some criterion looking for heterogeneity. Change point detection should seek to answer two important questions: “Did a change really occur?” and “When did the change occur?”. A major difficulty in change point detection is, however, the ability to detect changes that are not necessarily directly observed but are measured together with other types of disturbances.

Essentially change point detection algorithms can be classed as either parametric or non-parametric, with each being applied either retrospectively (*a posteriori*) or on-line (*a priori*). For this study both retrospective and on-line algorithms will be evaluated, complementing statistical data-based fault detection techniques by allowing the detection and analysis of potential interesting events that are occurring while the process is in a state of either normal or abnormal process operation.

One of the most basic change point detection techniques is based on the idea of splitting data into segments (2 or more), estimating some measure for each segment, and then determining how well the data fits the estimated measures. The mean square error (MSE) measure is one such estimator:

$$MSE(\phi) = \sum_{j=1}^{\phi} \left(X_j - \frac{\sum_{j=1}^{\phi} X_j}{\phi} \right)^2 + \sum_{j=\phi+1}^{n} \left(X_j - \frac{\sum_{j=\phi+1}^{n} X_j}{n-\phi} \right)^2 \qquad (4\text{-}1)$$

where X_j is data values, ϕ is the data change point separating the data segments and n is the total number of samples in the data set. The change point in a time series is determined by selecting ϕ in order to minimise $MSE(\phi)$. The procedure can then be repeated on the individual data segments. This method does not, however, guarantee that the identified change point is in fact real as in most cases a

Stellenbosch University http://scholar.sun.ac.za

minimised $MSE(\phi)$ can be obtained without any real change point existing in the time series. Various methods exist with which to overcome this limitation, with the techniques selected for evaluation being based on nearest-neighbours cumulative sums, Bayesian probability, singular spectrum analysis (SSA), extreme learning machine SSA and median significance testing. As a relatively basic, univariate change point detection technique, nearest-neighbours cumulative sums forms the basis of the change point detection analysis evaluation. Bayesian probability, although assuming input data to be independent and identically distributed, can be considered an improvement over nearest-neighbours cumulative sums, especially when analysing multivariate data of a known distribution. Singular spectrum analysis, being an extension of classical principal component analysis, is a non-parametric modelling technique that improves upon the Bayesian probability technique by not making any assumptions regarding the distribution of the data being analysed. Subsequently, extreme learning machine, having been shown to be good at generalisation and having extremely fast learning speeds with very few parameters that require setting (Huang et al., 2006), are considered as a non-linear operator used to remove non-linear characteristics from the data prior to SSA change point detection. Lastly, median significance testing, through the use of notched box plots, is performed as a visual confirmation of differences between data segments following change point detection analysis.

The techniques under consideration can be divided into procedural and interactive techniques. Whereas procedural techniques require only the specification of parameters and input data in order to arrive at usable results, interactive techniques require some inspection and possible adjustment of the results before any conclusions can be drawn.

This chapter deals with some of the theory behind change point detection and its application to process performance monitoring. The selected techniques are not an exhaustive list of available techniques, but have been chosen to be illustrative of what is available.

4.1 Procedural change point detection

4.1.1 Nearest-neighbours Cumulative Sums

Taylor (2000) showed that a combination of CUSUM charts and bagging can be used for change point detection. Firstly, cumulative sums (S_j) are calculated by adding the difference between the current value and the mean of the data to the previous sum (starting at zero). This gives the cumulative sums of differences between the values and the average, not the cumulative sums of the values. From this, the following can be deduced:

- a relatively straight segment of the CUSUM chart indicates a period where the values tend to be equal to the overall average
- an upward sloping segment of the CUSUM chart indicates a period where the values tend to be above the overall average
- a downward sloping segment of the CUSUM chart indicates a period where the values tend to be below the overall average

Stellenbosch University http://scholar.sun.ac.za

- a sudden change in direction in a segment of the CUSUM chart indicates a sudden change or shift in the data relative to the overall average

Unfortunately, identifying these sudden changes in direction in segments of the CUSUM chart is usually done very subjectively if performed manually. Using bagging, a confidence level can be determined for these apparent changes. This in turn, ensures that these sudden changes in direction are identified objectively. Taylor (2000) suggests using the difference between the maximum value of the CUSUM and the minimum value of the CUSUM (S_{diff}) as an estimator of the magnitude of the change. It has been found that this estimator works well regardless of the distribution of the data and despite multiple changes being present in the data.

Whereas bagging, being a distribution free approach, only assumes an independent error structure, control charting, based on the mean-shift model, includes assumptions such as that the data cannot be autoregressive and the random errors associated with the data needs to be identically distributed:

$$X_j = \mu_j + \varepsilon_j \qquad (4\text{-}2)$$

where X_j represents the data in time order, μ_j is the average of the data at time *j*, and ε_j is the random error associated with the *j*th value.

Bagging is performed by first randomly reordering the values in the data set (sampling without replacement). Next, the bagging CUSUM (S_j^0) of the randomly reordered values are calculated together with the bagging CUSUM difference (S_{diff}^0) and compared to the original CUSUM difference (S_{diff}). The results from multiple such bagging iterations (typically in excess of 1000) can then be used to estimate how much S_{diff} would vary if no change took place. By performing a large number of bagging iterations and calculating the percentage of bagging iterations for which S_{diff}^0 is less than S_{diff}, a confidence level that a change occurred can be calculated (typically for a significant change to be identified a confidence level of 90-95% is required). Furthermore, S_m, the point furthest from zero in the CUSUM chart or the maximum CUSUM deviation from the chord joining the CUSUMs at the first and last data points if only looking at a segment of the data set, can be used to identify the last point before the change occurred and S_{m+1} to identify the first point after the change occurred.

The span method (Woodward and Goldsmith, 1964) for identification of multiple change points can be summarised as:

1. A test is performed to determine if a change point exist in the data set.
2. If a change point exists, the significance of the maximum deviation is determined.

Stellenbosch University http://scholar.sun.ac.za

3. If the maximum deviation is significant, the change point is identified and the data set split into segments.
4. Next, the algorithm is repeated from step 1 for each data segment.

This method of change point detection is not limited to only the raw data value as input, but also other characteristics of the data such as mean, standard deviation, difference and rank (allowing smaller sustained changes to be detected with little interference from outliers), to name but a few. When using differences, it should however be noted that the differences cannot be calculated using common points as this will create correlation in the data.

Unfortunately, the span method gives an implied, but often unjustifiable, sequence of importance to the potential change points, where it is implied that the first identified change point is the most significant, and so on (Taylor et al., 2002). It does not explicitly consider which span should be used in assessing each potential change point. Typical examples of where this method fails are the "plateau" and "crossing the mean" problems. The "plateau" problem is characterised by a blip (possible outlier) in the CUSUM data set during a period of control between two real change points. The "crossing the mean" problem is characterised with the CUSUM data set crossing the mean while connecting two real change points. In both cases the span method could, and often do, easily incorrectly identify or miss the real change points. Human intervention is typically required to correct these failings of the span method.

An automated search method was suggested by Woodward and Goldsmith (1964) to correct for these failings:

1. Draw a cord from observation j, the current CUSUM value, back to the last change point or the first in the data set if no change point has been identified yet.
2. Calculate the maximum deviation between the CUSUM and the cord, identifying a potential change point, ϕ.
3. Test the potential change point, ϕ, for significance. If not significant, continue to next observation, $j+1$, and repeat from step 1 until a change point is identified or the end of the data is reached.
4. If the potential change point, ϕ, is significant, mark it as such and restart the search for the next change point at $\phi+1$.
5. Once completed the search is repeated backwards, accounting for effects of random variability, and the two sets of change points consolidated by amalgamating those sufficiently close.

This method is, however, far from ideal. Not only is the method more complex than the span method, it also requires extensive computational resources. Also, as before, all change points are given equal importance once identified and non-significant potential change points may mask real significant change points.

Stellenbosch University http://scholar.sun.ac.za

To address these issues, Taylor et al. (2002) developed the nearest-neighbours method that builds upon the span method. The key improvement of this method is the identification of three potential change points per iteration of the automated search method compared to the standard identification of one potential change point per iteration. The three potential change points can be selected using either of the following methods: (1) as the three points with the greatest absolute deviation from the cord or, (2) after finding the point with the maximum absolute deviation from the cord, the point with maximum absolute deviation from each of the two cords created are found (Figure 23). This method is still not perfect, and needs to be supported by review by a trained and experienced analyst.

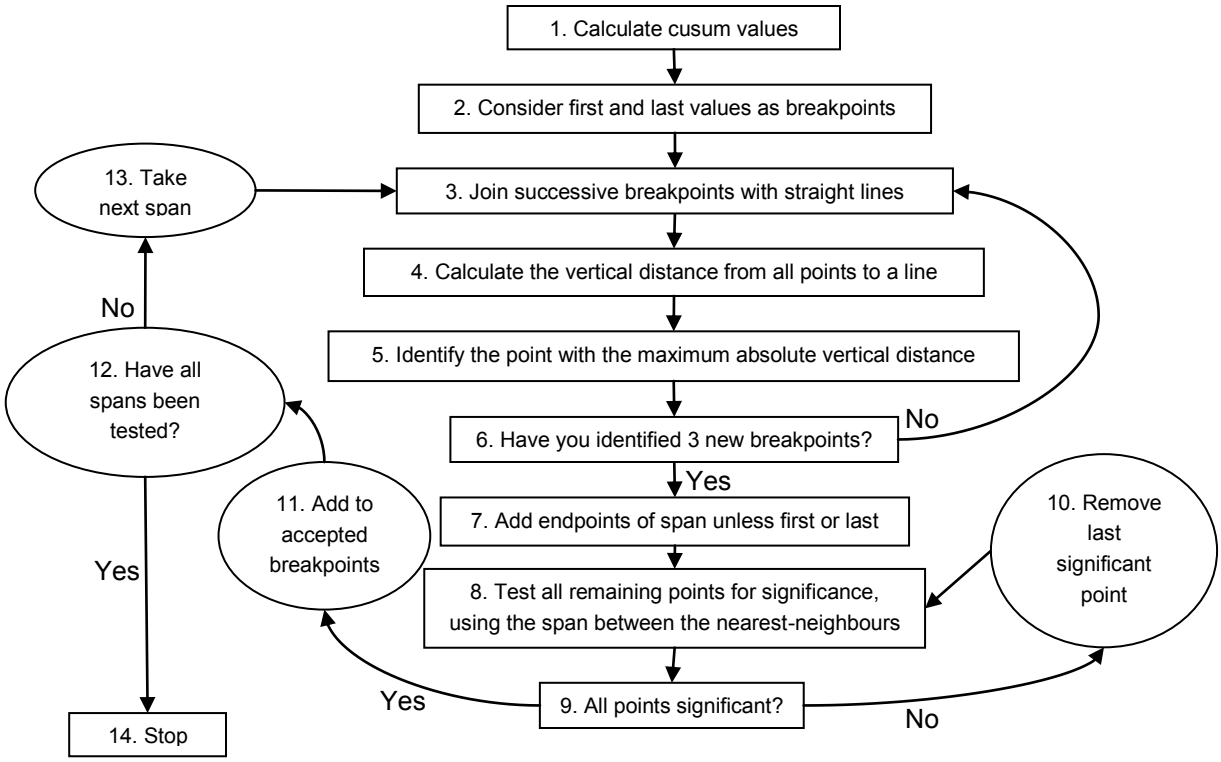

Figure 23: **Flowchart for nearest-neighbours method (adapted from Taylor et al., 2002)**

Given that change point analysis complements control charting, the technique based on cumulative sums and bagging does have the following advantages over control charting (Taylor, 2000):

- Better at detecting smaller sustained changes.
- Better at characterising changes (multiple changes and confidence levels).
- Reduces number of false detections by controlling the change-wise error rate.
- Robust to outliers.
- Very flexible.
- Simple to use and interpret.

Stellenbosch University http://scholar.sun.ac.za

On the negative side, however, the following shortcomings exist:

- Isolated abnormal points cannot be detected. Whereas the CUSUM (control chart) algorithm is optimal for detecting small shifts of the process mean, at the expense of being less sensitive to large changes, the Shewhart (control chart) algorithm is optimal for detecting isolated abnormal points.
- Due to the random selection of the bagging samples, identical, repeatable results cannot be obtained using the bagging approach. The precision of the results can be improved upon by increasing the number of bagging results.

4.1.2 Bayesian probability

Probability is typically used to describe the certainty that some event will occur and is expressed as a number between 0 and 1. The higher the probability of an event, the more certainty there is that the event will occur. Bayesian probability is probably the most popular version of subjective probability, using both expert knowledge and experimental data to produce probabilities. Whereas expert knowledge is represented by some prior probability distribution, expressing one's uncertainty about a quantity before the data is taken into account, the data is incorporated in a likelihood function. Bayes' theorem, a manipulation of conditional probabilities, multiplies the prior with the likelihood function, followed by normalisation, to produce the posterior probability distribution, the conditional distribution of the uncertain quantity given the data.

Considering that the joint probability of two events, ϕ_A & ϕ_B, can be expressed as:

$$\begin{aligned} P(\phi_A \phi_B) &= P(\phi_A \mid \phi_B) P(\phi_B) \\ &= P(\phi_B \mid \phi_A) P(\phi_A) \end{aligned} \tag{4-3}$$

Bayesian probability theory defines one of these events as the hypothesis, H, and the other as the data, X, where we wish to judge the relative truth of the hypothesis given the data (Olshausen, 2004). According to Bayes' theorem, this is accomplished via the relation:

$$P(H \mid X) = \frac{P(X \mid H) P(H)}{P(X)} \tag{4-4}$$

where $P(H \mid X)$ is the posterior, reflecting the probability of the hypothesis after consideration of the data, $P(X \mid H)$ is the likelihood function, assessing the probability of the observed data arising from the hypothesis, $P(H)$ is the prior, reflecting one's prior knowledge before the data is considered, and $P(X)$ the normalising constant. Usually the likelihood function is known as it expresses the knowledge of what the data is expected to look like given that the hypothesis is true. In contrast to this, the specification of the prior is probably the most subjective aspect of the theory.

Stellenbosch University http://scholar.sun.ac.za

Building on this foundation, Bayesian methods can easily be applied to detect changes in the generative parameters of time-series data. Generative models provide a way of modelling how a set of observed/evidence data could have arisen from a set of underlying causes. The most basic of these is probably a memory-less Bayesian generative change point model containing a built-in representation of change (Figure 24). For this model, Θ is a probability parameter, C represents whether or not a change has occurred, S represents the state responsible for generating the data, and X represents the observed/evidence data. Over time the state variable, S, form a Markov chain either taking on the value of the previous state at each iteration, if there is no change, or taking on some new value, if there is a change. What makes this model memory-less is the fact that the probability of a change occurring at time j is independent of all previous changes.

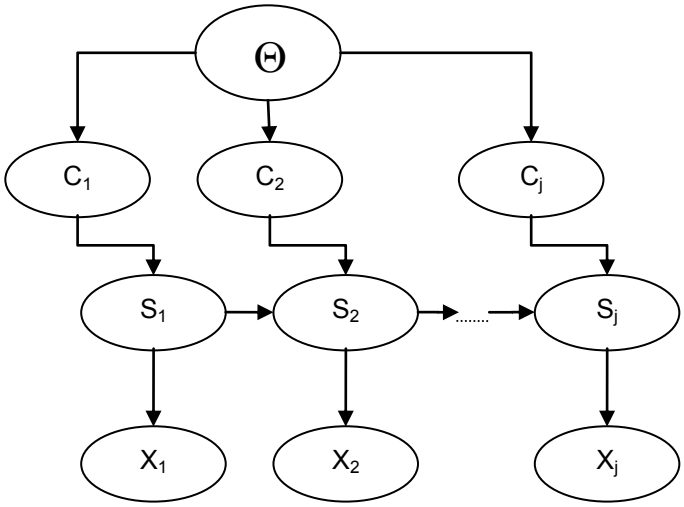

Figure 24: **Memory-less Bayesian generative change point model (adapted from Anderson, 2008)**

For online change point detection, Adams and MacKay (2006) cast the product partition model (Barry and Hartigan, 1992) into a Bayesian graphical model focussing on causal predictive filtering to generate an accurate distribution of the next unseen data in the sequence, given only data already observed. The product partition model assumes that time-series data can be separated based on changes in the data's generative parameters into partitions that are independent and identically distributed (iid). It is shown that if prior to making observations the probability distribution of random partitions is in a certain product form, given the observations it is also in product form. The model therefore provides a practical way for allowing the data to weight the partitions likely to hold, this being especially relevant to change point problems where the partitions divide the sequence of observations into components within which different regimes hold. By first conditioning on a partition and then averaging over all partitions an inference about particular future observations may subsequently be made. Finally, with suitable selection of prior product models, observations can ultimately approximate the true partition (Barry and Hartigan, 1992).

Stellenbosch University http://scholar.sun.ac.za

Given a sequence of observations $x_1, x_2, x_3, \ldots, x_n$ that may be divided into non-overlapping product partitions with the delineations between the partitions called the change points, the Bayesian online change point detection algorithm with prediction is defined as follow (Adams and MacKay, 2006):

1. Initialise:

   If some recent subset of the data is observed, the prior over the initial run length is the normalised survival function, $P(r_0) = \tilde{S}(r)$, alternatively all of the probability mass for the initial run length is placed at zero, $P(r_0 = 0) = 1$. Initialise the hyperparameters of the predictive distribution $\nu_1^{(0)} = \nu_{prior}$ and $\chi_1^{(0)} = \chi_{prior}$. Hyperparameters refer to the parameters of prior distributions, distinguishing them from the parameters of the model of the underlying data.

2. Observe new data:

   New data at time j is observed as x_j.

3. Evaluate predictive probability:

$$\pi_j^{(r)} = P\left(x_j \mid \nu_j^{(r)}, \chi_j^{(r)}\right) \tag{4-5}$$

4. Calculate growth probabilities:

$$P\left(r_{j-1} + 1, x_{1:j}\right) = P\left(r_{j-1}, x_{1:j-1}\right)\pi_j^{(r)}\left(1 - H\left(r_{j-1}\right)\right) \tag{4-6}$$

5. Calculate change point probabilities:

$$P\left(r_j = 0, x_{1:j}\right) = \sum_{r_{j-1}} P\left(r_{j-1}, x_{1:j-1}\right)\pi_j^{(r)} H\left(r_{j-1}\right) \tag{4-7}$$

6. Calculate evidence:

$$P\left(x_{1:j}\right) = \sum_{r_j} P\left(r_j, x_{1:j}\right) \tag{4-8}$$

7. Determine run length distribution:

Stellenbosch University http://scholar.sun.ac.za

Since the predictive distribution depends only on the recent data $x_j^{(r)}$, the posterior distribution, $P(r_j \mid x_{1:j}) = \frac{P(r_j, x_{1:j})}{P(x_{1:j})}$, is thus found by writing the joint distribution over the run length and observed data recursively.

8. Update statistics:

$$\begin{aligned} \nu_{j+1}^{(0)} &= \nu_{prior} \\ \chi_{j+1}^{(0)} &= \chi_{prior} \\ \nu_{j+1}^{(r+1)} &= \nu_j^{(r)} + 1 \\ \chi_{j+1}^{(r+1)} &= \chi_j^{(r)} + u(x_t) \end{aligned} \qquad (4\text{-}9)$$

9. Perform prediction:

The marginal predictive distribution is calculated by integrating over the posterior distribution on the current run length, $P(x_{j+1} \mid x_{1:j}) = \sum_{r_j} P(x_{j+1} \mid x_j^{(r)}, r_j) P(r_j \mid x_{1:j})$.

10. Return to Step 2

The algorithm is concerned with estimating the posterior distribution over the current run length, or time since the last change point, given the data so far observed and ensuring the algorithm gradually forgets the effect of past data (Yamanishi and Takeuchi, 2002). It is assumed that for each partition, ρ, the data within it are iid from some probability distribution $P(x_j \mid \eta_\rho)$ with the parameters $\eta_\rho, \rho = 1,2,3,\ldots$ also taken to be iid. The length of the current run at time j is denoted by r_j, with $x_j^{(r)}$ denoting the set of observations associated with the run. The run length can be interpreted as the length of time (or equivalently, the number of observations) since the last change point was observed. A low posterior, $P(r_j \mid x_{1:j})$, is therefore indicative of a change in the data characteristics at time j, and consequently a change in the process on which the data are measured. Since the posterior over possible partitions is a marginal density coming from an average over all possible generative parameters, the choice of prior hyperparameters in the algorithm is crucial in its ability to successfully detect change points.

Due to the potential of miscalled choices of the prior hyperparameter settings, Paquet (2007) proposed an empirical Bayesian treatment of the prior hyperparameters used in the Bayesian online change point detection algorithm, by relying on the standard addition of a backwards loop to the algorithm. First, a lower bound is constructed on the marginal density, giving a practical handle on the log marginal

Stellenbosch University http://scholar.sun.ac.za

likelihood. Subsequently, an expectation maximisation algorithm is used to maximise the log marginal likelihood, with a variational treatment of latent variables being applied in the E-step of the algorithm and a concave-convex procedure being applied in the M-step of the algorithm. This addition to the algorithm does, however, introduce a constraint in the form of a restriction of the generative probabilities to the exponential family of models with conjugate priors.

4.1.3 Singular Spectrum Analysis

Singular spectrum analysis (SSA), being an extension of classical principal component analysis, is another method than can be applied to analyse and detect a change in time series. In contrast to the nearest-neighbours CUSUM method, SSA allows for the time series to be of complex structure. SSA can be viewed as a linear root mean square fitting method (Kugiumtzis and Christophersen, 1997) where all available samples in a time window length, τ_N, is processed with singular value decomposition (SVD), resulting in the final transformation being linear combinations of the samples: separating the time series into signal and noise components.

In general, SSA is the application of singular value decomposition of the trajectory matrix obtained from the original time series with a subsequent reconstruction of the time series. The main concept in studying the properties of SSA is “separability”, which characterises how well different components can be separated from each other. The SSA method works with arbitrary statistical processes whether linear or non-linear, stationary or non-stationary, Gaussian or non-Gaussian, making no prior statistical assumptions about the data such as stationarity of the series or normality of the residuals. The basic SSA algorithm decomposes a data set into its component parts and reconstructs the data set by leaving the random (noise) component behind (Moskvina and Zhigljavsky, 2003):

1. Embedding:

   A large state vector is first derived from successive samples from a time series, $x_1, x_2, x_3, \ldots, x_N$, by embedding with a delay of unity to form a trajectory matrix:

$$X = \left(x_{ij}\right)_{ij=1}^{M,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & x_{M+2} & \cdots & x_N \end{pmatrix} \tag{4-10}$$

   where M ($M \leq N/2$) is a lag parameter and $K = N - M + 1$. X, having equal elements on the diagonals, is also referred to as a Hankel matrix. X can further be considered as a multivariate data with M characteristics and K observations where $R = XX^T$ is called the lag-covariance matrix.

Stellenbosch University http://scholar.sun.ac.za

2. Singular value decomposition:

   Singular value decomposition (SVD) is used to decompose the lag-covariance matrix, providing a collection of M eigenvectors (chosen to accommodate the most populated directions), eigenvalues (relate the degree to which the data fill the new directions) and principal components.

   Given the eigenvalues of R, $\lambda_1, \lambda_2, \ldots, \lambda_M$, arranged in decreasing order, $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_M \geq 0$, the corresponding orthonormal eigenvectors of R, $U_1, U_2, \ldots, U_M$, the number of nonzero eigenvalues, d, and V_i, the eigenvectors of the matrix $X^T X$, we have:

   $$V_i = X^T U_i \qquad (4\text{-}11)$$

   For $i = 1,2,\ldots,d$. Through SVD X is represented as $X = X_1 + X_2 + \ldots + X_d$, where $X_i = \sqrt{\lambda_i} U_i V_i^T$ are rank-one biorthogonal matrices.

3. Grouping:

   The set of indices $\{1,2,\ldots,d\}$ is split into two groups, $I = \{i_1, i_2, \ldots, i_l\}$ and $\bar{I} = \{1,2,\ldots,d\} \setminus I$, and the matrices X_i within each group is summed. X is now represented as:

   $$X = X_I + X_{\bar{I}} \qquad (4\text{-}12)$$

   where $X_I = \sum_{i \in I} X_i$ and $X_{\bar{I}} = \sum_{i \notin I} X_i$.

4. Reconstruction:

   First, averaging is performed over the diagonals $i + j = const$ of the matrices X_I and $X_{\bar{I}}$. Next, the one-to-one correspondence between the series of length N and the Hankel matrices of size $M \times K$ is applied twice, resulting in two series and the SSA decomposition of the original series (x_j):

   $$x_j = z_j + e_j \qquad (4\text{-}13)$$

Stellenbosch University http://scholar.sun.ac.za

where z_j can often be associated with signal and e_j (the residual series) with noise.

For change point detection it is expected that if at a certain time moment $N+\tau$ the mechanism generating x_j $\left(j \geq N+\tau\right)$ has changed, the distance between the p-dimensional subspace and lagged vectors X_j for $j \geq K+\tau$ would have increased. This assumption translates into detection of conditions resulting from increased variance of the system and, depending on the nature of the system, changes in dynamic topology may also be detected. Rather than applying SVD to the standard SSA algorithm trajectory matrix, the trajectory matrix computed in a time interval $\left[n+1, n+N\right]$ of length N is used, not only making the algorithm sequential but also adapts it to slow change, multiple changes and outliers. For a time series $x_1, x_2, x_3, \ldots$ and fixed integers N, M, p, w and q where $p < M \leq N/2$ and $0 \leq w < q$, Moskvina and Zhigljavsky (2003) define the SSA change point detection algorithm as follows (executed for each $n = 0,1,2,\ldots$):

1. p-Dimensional space construction:

   Perform the first three steps of the SSA algorithm in the time interval $\left[n+1, n+N\right]$:

   - Construct the trajectory matrix:

$$X_B^{(n)} = \begin{pmatrix} x_{n+1} & x_{n+2} & x_{n+3} & \cdots & x_{n+K} \\ x_{n+2} & x_{n+3} & x_{n+4} & \cdots & x_{n+K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & x_{n+M+2} & \cdots & x_{n+N} \end{pmatrix} \qquad \text{(4-14)}$$

   - Perform the SVD of the lag-covariance matrix $R_n = X^{(n)}\left(X^{(n)}\right)^T$, giving a collection of M eigenvectors.

   - Select a particular group I of $p < M$ of these eigenvectors.

2. Test matrix construction:

   Construct the test matrix $X_T^{(n)}$ of size $M \times Q$:

Stellenbosch University http://scholar.sun.ac.za

$$X_T^{(n)} = \begin{pmatrix} x_{n+w+1} & x_{n+w+2} & x_{n+w+3} & \cdots & x_{n+q} \\ x_{n+w+2} & x_{n+w+3} & x_{n+w+4} & \cdots & x_{n+q+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+w+M} & x_{n+w+M+1} & x_{n+w+M+2} & \cdots & x_{n+w+M-1} \end{pmatrix} \quad (4\text{-}15)$$

where $q = w + Q$.

3. Detection statistics computation:

Three detection statistics exist for determining change points:

- The sum of squared Euclidean distances between the vectors $X_i^{(n)}(i = w+1, \ldots, q)$ and the p-dimensional subspace of $\Re^M$:

$$D_{n,I,w,q} = \sum_{i=p+1}^{q} \left(\left(X_i^{(n)}\right)^T X_i^{(n)} - \left(X_i^{(n)}\right)^T U U^T X_i^{(n)} \right) \quad (4\text{-}16)$$

- The normalised sum of squares of distances:

$$S_n = \frac{D_{n,I,w,q}}{MQ\mu_{n,I}} \quad (4\text{-}17)$$

where $\mu_{n,I} = \tilde{D}_{m,I,0,K}$ (where m is the largest value of $m \le n$ so that the hypothesis of no change is accepted) is an estimator of the normalised sum of squared distances $\tilde{D}_{n,I,w,q} = \frac{D_{n,I,w,q}}{MQ}$ at the time intervals $[j+1, j+m]$, also where the hypothesis of no change can be accepted.

- The CUSUM-type statistic:

$$W_1 = S_1,\ W_{n+1} = \max\left\{0, \left(W_n + S_{n+1} - S_n - \kappa / \sqrt{MQ}\right)\right\},\ n \ge 1 \quad (4\text{-}18)$$

where κ is a small nonnegative constant, with a reasonable value being $\kappa = 1/\left(3\sqrt{MQ}\right)$ (Moskvina, 2001).

Stellenbosch University http://scholar.sun.ac.za

Typically, large values of $D_{n,I,w,q}$, S_n and W_n will indicate structural change in the time series. More specifically, structural change is said to have occurred in the time series for some n if $W_n > h$:

$$h = \frac{2t_\alpha}{MQ}\sqrt{\frac{1}{3}Q\left(3MQ - Q^2 + 1\right)} \qquad (4\text{-}19)$$

where t_α is the $(1-\alpha)$ quantile of the standard normal distribution.

Significant structural changes in a time series will be detected for any reasonable choice of parameters. However, tuning of the parameters may be required to detect small changes in noisy time series. Reasonable detection parameters can be defined as follow:

- With the SSA algorithm the choice of embedding dimension, K, is not critical and any choice over a lower limit will give essentially the same reconstruction because the additional coordinates correspond to less significant singular values and give negligible variance assuming τ_N is sufficiently large (Kugiumtzis and Christophersen, 1997).
- For the SSA algorithm achieving "independence" or "separability" of the components z_t and e_t in the SSA decomposition is of key importance. If N is not very large it is recommended to choose the lag, M, as $M = N/2$ and the group, I, as $I = \{1,\ldots,p\}$ where l is such that the first p components provide a good description of the signal and the lower $M - p$ components correspond to noise (Moskvina and Zhigljavsky, 2003). To choose p the SSA decomposition of the whole series and some large part of the series should be visually inspected prior to applying the change-point detection algorithm. Choosing p too small may result in underfitting, causing parts of the signal, and potential changes, to be missed. Choosing p too large may result in overfitting, causing parts of noise to be approximated together with the signal, making finding a true change in the signal potentially more difficult.
- For the length of the test sample, w, it is recommended to choose $w \geq K$, making the algorithm more sensitive to changes than its more economical version when $w < K$. For the location of the test sample, q, it is recommended to select q slightly larger than w to ensure smooth behaviour from the test statistics $D_{n,I,w,q}$. If the difference between w and q becomes too large, the behaviour of $D_{n,I,w,q}$ becomes too smooth.
- For the window width, N, it is recommended to choose a reasonably large value of N. If N is too small, outliers may be recognised as structural changes in the time series. If N is too large, the possibility exist that all changes in the time series may be missed or smoothed out. It must also be noted that if small gradual changes in the time series is allowed, N may not be very large.

- If either a constant or a linear component whose influence should be neglected is contained in the trend of the original series $\{x_t\}$, centering of the base and test matrices may be worthwhile. Centering consists of subtracting row (and possibly column) averages from the elements of the matrices.

Optimisation of the SSA parameters is a simple way of improving the change point detection capability of the algorithm. As mentioned, the choice of embedding dimension, K, is not critical, however, optimising the lag parameter, M, the point of maximum decorrelation taken over all variables, may improve results. The point of decorrelation is taken as the time delay, M. If M is selected too small, the resulting vectors may be very nearly the same, each carrying a great deal of redundant information, while too large a M may produce coordinates that are essentially unrelated. Small errors in the data will also become exponentially magnified in time, resulting in a too large a delay and decorrelating the signal from itself (Abarbanel, 1996). The correct delay, maximising the independence between the first two state vectors, should therefore be used.

A widely accepted criteria for the selection of M is that the components of the vector x_i must be uncorrelated (Kugiumtzis, 1996). To meet this criteria, the estimates of M are based either on linear decorrelation, using the linear autocorrelation function, or general decorrelation, using average mutual information (AMI). Although other methods are also available for delay time estimation, these methods are most commonly used, with AMI having been suggested to be the most robust in the presence of noisy data (Abarbanel, 1996).

By measuring the degree of correlation of a variable at one time with itself at another time, the autocorrelation function determines to what extent one part of the time series looks like another part of the same time series. For computational purposes, the autocorrelation is defined as follow:

$$\begin{aligned} autocorrelation &= \frac{autocovariance}{variance} \\ &= \frac{\sum_{j=1}^{N-M}\left(x_j-\bar{x}\right)\left(x_{j+k}-\bar{x}\right)}{\sum_{j=1}^{N}\left(x_j-\bar{x}\right)^2} \end{aligned} \qquad (4\text{-}20)$$

Mutual information, similar to the autocorrelation function, tries to measure the extent to which values of $x_{\tau+M}$ are related to values of x_τ, at a given lag. Rather than using a linear basis to assess the correlation, mutual information has the advantage of using probabilities (Nichols and Nichols, 2001):

$$AMI = \sum_i px_i \log_2\left(px_i\right) + \sum_j py_j \log_2\left(py_j\right) - \sum_i \sum_j px_i y_j \log_2\left(px_i y_j\right) \qquad (4\text{-}21)$$

Stellenbosch University http://scholar.sun.ac.za

The AMI statistics relies on the fact that when coordinates of a point have the same value, they represent the same information (maximum mutual information) and one can be used to exactly predict the other. For reconstruction, x_τ should not provide a lot of information about $x_{\tau+M}$ (minimum mutual information), with mutual information, at a given lag, being relatively high for strongly related and relatively low for weakly related values (Williams, 1999). Therefore, to determine the point at which the values become independent from one another, the first minimum of the AMI plot for a time series is used. Not only having a low AMI value, but also having a distribution of mutual information which is sharp, is important in that a single choice of M can accurately capture the underlying dynamical process (Abarbanel et al., 2001).

With the autocorrelation function expressing a linear relationship between the signal and itself at a given delay, the AMI statistic, not being tied to linear/non-linear properties of the system (Abarbanel, 1996), represents a clear improvement over the autocorrelation function (Nichols and Nichols, 2001) and is generally favoured for delay time estimation when analysing non-linear systems. Very large data sets are, however, required when using the AMI statistic for delay time estimation, rendering many data points unusable (Atmanspacher et al., 1988). Furthermore, even with both methods guaranteeing decorrelation between two successive components x_i and x_{i+M} of the reconstructed vector x_i, even if x_i and x_{i+M} are uncorrelated and x_{i+M} and x_{i+2M} are uncorrelated, it cannot be guaranteed that that x_i and x_{i+2M} would also be uncorrelated (Kugiumtzis, 1996).

Considering these disadvantages, together with the fact that the AMI statistic may not have a clear minimum or the autocorrelation function may require an extremely long time to reach approximately zero, it is evident that the analysis of non-linear systems is not algorithmic alone, but also requires some experience and intuition. Continuous monitoring of processes is, however, still possible following off-line parameter estimation and model validation.

4.1.4 Extreme learning machine SSA

As with the statistical data-based fault detection, extreme learning machine provides an opportunity with which to improve the singular spectrum analysis change point detection algorithm (Moskvina and Zhigljavsky, 2003). Similar to the ELM PCA fault detection technique, ELM is used to construct non-linear models that are used for residual generation, prior to residual evaluation by SSA. For NOC data, the residuals are expected to be closed to zero, while under abnormal operating conditions, the zero point of the residual variable would drift away, allowing for change point detection. The ELM therefore acts as a non-linear dynamic operator used to remove the non-linear and dynamic characteristics, prior to the application of SSA. For the sake of completeness, the overview of ELM from section 3.4.4 is restated here.

ELM is a learning algorithm for SLFNs where the input weights are chosen randomly and the output weights determined analytically (Huang et al., 2006). Whereas traditional learning algorithms employed

Stellenbosch University http://scholar.sun.ac.za

by feedforward neural networks are limited in their learning speed, mainly due to the slow gradient descent-based nature of the algorithms and the iterative parameter estimation, ELM can not only achieve extremely fast learning speed but also tends to provide good generalization performance (thanks to the learning algorithm not only typically reaching the smallest training error but also the smallest norm of weights). Furthermore, for conventional feedforward neural networks there exists a dependency between the different layers of weight and bias parameters which requires all parameters to be tuned. For SLFNs with N hidden nodes it has been shown that with randomly chosen input weights and hidden layer biases exactly N distinct observations can be learned (Huang, 2003), limiting the aforementioned dependency and significantly reducing the number of parameters that needs to be estimated. This allows such an SLFN to be considered as a linear system with the output weights being analytically determined through simple generalized inverse operation of the hidden layer output matrices (Huang et al., 2006), also making it ideal for use when analyzing very large data sets.

Given a training data set consisting of N arbitrary distinct samples (x_j, y_j), a hidden node number $\tilde{N}$, and an activation function $g(x)$, the ELM algorithm can be defined as (Huang et al., 2006):

1. Randomly assign the input weights z_j and biases b_j, $j = 1, \ldots, \tilde{N}$.
2. Calculate the hidden layer output matrix P.
3. Calculate the output weights $\beta = P^H Y$ where $Y = [y_1, \ldots, y_N]^T$

As with the SSA change point detection algorithm (Moskvina and Zhigljavsky, 2003), for change point detection it is expected that if at a certain time moment $N + \tau$ the mechanism generating x_j $(j \geq N + \tau)$ has changed, the distance between the p-dimensional subspace and lagged vectors X_j for $j \geq K + \tau$ would have increased. This assumption translates into detection of conditions resulting from increased variance of the system and, depending on the nature of the system, changes in dynamic topology may also be detected. However, for the ELM SSA change point detection algorithm both the trajectory and test matrices are first converted to residual matrices through the application of the ELM algorithm prior to SVD.

For a time series $x_1, x_2, x_3, \ldots$ and fixed integers N, M, p, w and q where $p < M \leq N/2$ and $0 \leq w < q$, the ELM change point detection algorithm can be defined as follows (executed for each $n = 0,1,2,\ldots$):

1. p-Dimensional space construction:

   Perform the first three steps of the SSA algorithm in the time interval $[n+1, n+N]$:

Stellenbosch University http://scholar.sun.ac.za

- Construct the trajectory matrix:

$$X_B^{(n)} = \begin{pmatrix} x_{n+1} & x_{n+2} & x_{n+3} & \cdots & x_{n+K} \\ x_{n+2} & x_{n+3} & x_{n+4} & \cdots & x_{n+K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & x_{n+M+2} & \cdots & x_{n+N} \end{pmatrix} \quad (4\text{-}22)$$

- Randomly assign the ELM input weights z_i and biases b_i, $i = 1, \ldots, \tilde{N}$.

- Calculate the ELM hidden layer output matrix $P_B^{(n)}$.

- Calculate the ELM output weights $\beta = P^H Y$

- Calculate the residual trajectory matrix, $X_B^{(n)}$, as the difference between the trajectory matrix and the predicted ELM outputs.

- Perform the SVD of the residual-covariance matrix $R_n = X^{(n)}\left(X^{(n)}\right)^T$, giving a collection of M eigenvectors.

- Select a particular group I of $p < M$ of these eigenvectors.

2. Test matrix construction:

- Construct the test matrix $X_T^{(n)}$ of size $M \times Q$:

$$X_T^{(n)} = \begin{pmatrix} x_{n+w+1} & x_{n+w+2} & x_{n+w+3} & \cdots & x_{n+q} \\ x_{n+w+2} & x_{n+w+3} & x_{n+w+4} & \cdots & x_{n+q+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+w+M} & x_{n+w+M+1} & x_{n+w+M+2} & \cdots & x_{n+w+M-1} \end{pmatrix} \quad (4\text{-}23)$$

where $q = w + Q$.

Stellenbosch University http://scholar.sun.ac.za

- Calculate the residual test matrix, $X_T^{(n)}$, as the difference between the test matrix and the predicted ELM outputs.

3. Detection statistics computation:

Three detection statistics exist for determining change points:

- The sum of squared Euclidean distances between the vectors $X_i^{(n)}(i = w+1,\ldots,q)$ and the p-dimensional subspace of $\Re^M$:

$$D_{n,I,w,q} = \sum_{i=p+1}^{q} \left(\left(X_i^{(n)}\right)^T X_i^{(n)} - \left(X_i^{(n)}\right)^T U U^T X_j^{(n)} \right) \qquad (4\text{-}24)$$

- The normalised sum of squares of distances:

$$S_n = \frac{D_{n,I,w,q}}{MQ\mu_{n,I}} \qquad (4\text{-}25)$$

where $\mu_{n,I} = \tilde{D}_{m,I,0,K}$ (where m is the largest value of $m \le n$ so that the hypothesis of no change is accepted) is an estimator of the normalised sum of squared distances $\tilde{D}_{n,I,w,q} = \frac{D_{n,I,w,q}}{MQ}$ at the time intervals $[j+1, j+m]$, also where the hypothesis of no change can be accepted.

- The CUSUM-type statistic:

$$W_1 = S_1,\ W_{n+1} = \max\left\{0, \left(W_n + S_{n+1} - S_n - \kappa / \sqrt{MQ}\right)\right\},\ n \ge 1 \qquad (4\text{-}26)$$

where κ is a small nonnegative constant, with a reasonable value being $\kappa = 1/\left(3\sqrt{MQ}\right)$ (Moskvina, 2001).

Stellenbosch University http://scholar.sun.ac.za

4.2 *Interactive change point detection*

4.2.1 Median significance

Although it is unavoidable for change point detection techniques to signal some false change points, most often human interaction with the results can assist with distinguishing between true and false change points. Typical things looked at include the relative frequency of change points identified as well as the magnitude of the measure that identified the change points. Graphical inspection of the raw data with or without procedurally determined change point detection results allows one to identify new suspected change points in the data or validate change points detected through the procedural techniques. The data can then be visually segmented into portions where no significant changes are considered to have occurred and significance tests performed on these successive data portions. The significance tests can determine if statistically significant differences exist in the distribution parameters of each of the portions. One such significance test is the notched box plot for inspecting significant shifts in the median of successive data portions (Figure 25).

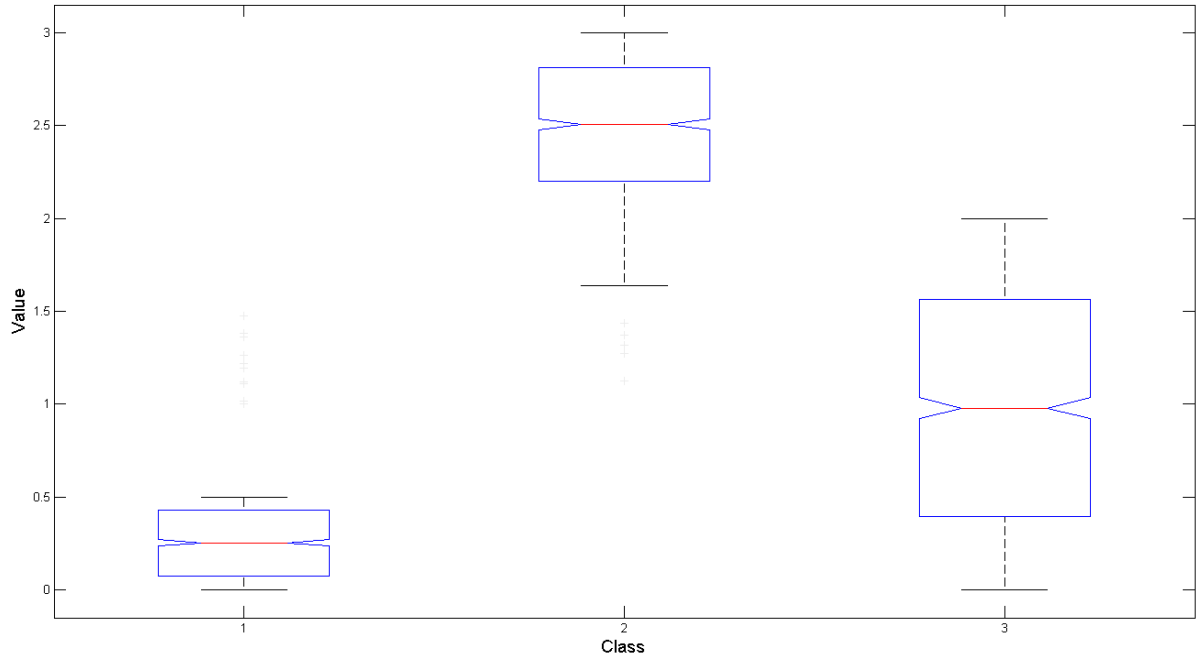

Figure 25: **A notched box plot for inspecting significant shifts in the median of successive data portions**

For multivariate data, the dimensionality of the data can be reduced using any of the PCA type techniques, and relevant statistics displayed for change point detection. Two of the most basic to use

Stellenbosch University http://scholar.sun.ac.za

include the Hotelling's T^2 statistic and the Q statistic (SPE) which can be used to view multiple variables as a single time series.

For the notch box plots, [0.25 0.5 0.75] quantiles are generated for each user or procedurally determined data portion of either an individual variable, a change point detection statistic or one of the multivariate statistics (Hotelling's T^2 statistic or the Q statistic). Notches are also used to display the 95% significance level for changes in the median of consecutive data portions. Although the significance level is based on the normal distribution assumption, the comparison of medians is fairly robust for other distributions. This comparison of medians using notched box plots is the visual hypothesis test analogue to the t-test for means.

4.3 Technique evaluation

For the evaluation of the change point detection techniques, the 3 data sets / models as described in section 3.5 were used:

- Simple multivariate time series data
- Simple multivariate process (Ku et al., 1995)
- Tennessee Eastman process (Downs and Vogel, 1993)

The evaluation not only focusses on validation of the techniques, but also allows for an appreciation of the techniques and their applicability to be obtained and a comparison to be drawn with the results achieved by the statistical data-based fault detection techniques.

The change point detection metrics were subsequently tested on both the NOC data as well as all the predetermined fault condition data for each of these data set / models. Listed in Table 3 is an overview of all the performance metrics tested.

Table 3: **Overview of change point detection techniques**

Technique	Type	Metric	Comments
Nearest-neighbours CUSUM	Linear	Maximum deviation	Complex, univariate, computationally expensive technique.
Bayesian probability	Linear	Probability	Assume data to be independent and identically distributed. Can accommodate data from different distributions.
Singular spectrum analysis	Linear	S_n W	Does not require data to be independent and identically distributed. Assume data to be normally distributed. Exploit autocorrelation in data.
Extreme learning machine SSA	Non-linear	S_n W	No assumption regarding data distribution. Extremely fast learning speed. Collinearity potentially a problem.

Stellenbosch University http://scholar.sun.ac.za

As required by the Bayesian probability, SSA and ELM SSA techniques, data were scaled to zero mean and unit variance prior to these analyses. Furthermore, prior to evaluation the variables for each data set were checked to see whether or not the data are independent and identically distributed and have a normal distribution. To test for iid data, each variable was checked to determine if there was any significant autocorrelation in the time series. Significant levels of autocorrelation are indicative of data that is not independent and identically distributed. To test if the data is normally distributed, each variable was subjected to the Lilliefors test. The Lilliefors test is a goodness-of-fit test of composite normality, testing whether or not the data in a time series come from an unspecified normal distribution. Such tests to determine the statistical characteristics of the data are critical in that they allow appropriate techniques, having specific assumptions, to be matched up with the data.

For the evaluation of the various change point detection techniques, the results were both objectively and subjectively considered. A peak finding algorithm was applied to the results to automatically identify detected change points. The change points identified through the peak finding algorithm for the individual runs (vertical dotted blue lines on the change point detection graphs) were subsequently combined over all runs (vertical light blue bars on the change point detection graphs) to emphasize recurring change points, henceforth referred to as confirmed peaks.

For the SSA and ELM SSA change point detection 2 statistics are calculated: the normalised sum of squares of distances S_n, and the CUSUM-type statistic, W. Of these, significance limits of 95% and 99% (horizontal dotted lines on the change point detection graphs) have been calculated for W to confirm change points based on the magnitude of the statistic. For S_n visual inspection of the statistic in combination with the confirmed peaks indicator was used to confirm change points. For the Bayesian change point detection, a 99% significance limit was introduced for the probability estimation. As with some of the SSA and ELM SSA change point detection statistics, visual inspection of the statistic in combination with the confirmed peaks indicator was used to confirm change points. Unlike the multivariate SSA, ELM SSA and Bayesian change point detection algorithms, the nearest-neighbours CUSUM change point detection algorithm is a univariate technique, detecting change points in individual variables. For the nearest-neighbours CUSUM change point detection, the detected change points in individual variables where combined in the form of confirmed peaks, the results of which were subsequently used to calculate the Hotelling's T^2 statistic for the combined data set. A combination of visual inspection of the statistics and the confirmed peaks indicator was used to confirm change points.

4.3.1 Simple multivariate time series data

For the simple multivariate time series data 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 501, indicated by the black vertical line on the change point detection evaluation figures, and the evaluation criteria determined over the following 500 data points. For determining the reference models the average of 10

Stellenbosch University http://scholar.sun.ac.za

data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. Where required, a confidence limit threshold of 99% was used.

As stated previously, inspection of the simple multivariate time series data (Figure 10) indicated it to be independent and identically distributed while having a normal distribution. These data characteristics meet the required assumptions of all of the change point detection techniques; ensuring their relevance to the data set being analysed. Furthermore, since the data set comprises entirely of randomly generated data with only a predefined covariance between x_2, x_3 and x_4 it is expected that, although applicable, there would be no need for the additional complexity offered by non-linear techniques. For the sake of completeness, however, all the available algorithms will be assessed. This will not only allow all appropriate algorithms to be evaluated but also help to gain insight into the potential benefits that may be offered by the more complex, non-linear change point detection techniques.

From the analysis of the NOC data (Table 4), no false events were detected for the nearest-neighbours CUSUM, the Bayesian probability, SSA W or the ELM W statistics. For the SSA S_n and ELM S_n statistics, some false events were detected but easily rejected following visual inspection due to their small magnitude and frequency of occurrence.

Table 4: **Simple multivariate time series data: NOC results**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
MTS0	No events	No events	No events	No events	No events	No events

For the step change fault conditions, MTS1-MTS4, all the techniques were able to easily detect the event in the data set (Table 5). Visual inspection of the various change point detection statistics confirmed this (Figure 26).

For the ramp change fault conditions, MTS5-MTS8, none of the techniques were able to detect the event in the data set (Table 5). Inspection of the statistics (Figure 27) shows that although no clear indication is given, it is possible to visually identify the event for the nearest-neighbours CUSUM T^2, the Bayesian probability and the ELM S_n statistics, although after a slight delay. It is therefore evident that due to the nature of the ramp change and the fact that all of the techniques are looking for more pronounced changes in the data, none of the change point detection techniques are effective in detecting a slow drift in the data.

For the spike fault conditions, MTS9-MTS12, all the techniques were able to easily detect the event in the data set (Table 5). Unfortunately, none of the techniques were able to detect the second change point in

Stellenbosch University http://scholar.sun.ac.za

the data for the spike having a longer duration in the uncorrelated variable (Figure 28), with only the nearest-neighbours CUSUM T^2 and Bayesian Probability statics being able to detect this in the correlated variable. The fact that none of the SSA or ELM SSA statistics were able to detect the second event can be attributed to the fact that the SSA technique embeds the data, effectively passing a moving window filter over the data. This combined with the relatively short duration of the spike, causes the SSA and ELM SSA change point detection techniques to be ineffective.

For the standard deviation change fault conditions, MTS13-MTS16, none of the techniques were able to detect the event in the data set (Table 5). Visual inspection of the various change point detection statistics confirmed this (Figure 29).

For the covariance shift fault conditions, MTS17-MTS19, only the Bayesian probability statistic was able to detect all change points (Table 5), although, due to the magnitude of the Bayesian probability, confirmation was needed via visual inspection for the covariance shift in a single variable. The nearest-neighbours CUSUM statistic was not able to detect any of the change points, the nearest-neighbours CUSUM statistic being a univariate technique and the covariance shift being a multivariate change, with the SSA and ELM SSA statistics being able to only detect the most significant variance shift (Figure 30).

For the function change fault conditions, MTS20-MTS23, again only the Bayesian probability statistic was able to detect all change points (Table 5), although, due to the magnitude of the Bayesian probability, confirmation was again needed via visual inspection for the smaller changes. The SSA and ELM SSA statistics were only able to partly detect the function change fault conditions, with visual inspection of the nearest-neighbours CUSUM statistics showing that although the technique was unable to detect any change points, with visual inspection the larger function change fault conditions could be observed via the multivariate representation of the data (Figure 31).

It can be concluded that for the simple multivariate time series data, that is independent and identically distributed with a normal distribution, as expected the Bayesian change point detection technique performed the best of the techniques tested, correctly detecting 78.3% of the change points. The technique was able to successfully detect most of the fault conditions outright, with only a few requiring visual inspection as confirmation, and only the standard deviation fault conditions not being detectable. The ELM SSA technique was able to correctly detect 73.9% of the change points, being better than the SSA technique at detecting more subtle changes in the data such as the ramp change (drift). Whereas the nearest-neighbours CUSUM technique struggled specifically with the multivariate changes to the data set, only being able to correctly detect 56.5% of the change points, the SSA technique correctly detected 65.2% of the change points, its performance hampered by the fact that the data was independent and identically distributed. This is a direct result from the fact that the SSA technique exploits the collinearity in the lag trajectory matrix to achieve optimal signal separation, requiring some autocorrelation in the data.

Stellenbosch University http://scholar.sun.ac.za

Table 5: **Simple multivariate time series data: fault condition results (Yes = correct change point clearly identified; Graph = correct change point visually inferred; No = correct change point not identified)**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
MTS1	Yes	Yes	Yes	Yes	Yes	Yes
MTS2	Yes	Yes	Yes	Yes	Yes	Yes
MTS3	Yes	Yes	Yes	Yes	Yes	Yes
MTS4	Yes	Yes	Yes	Yes	Yes	Yes
MTS5	Graph	Graph	No	No	Graph	No
MTS6	Graph	Graph	Graph	No	Graph	No
MTS7	Graph	Graph	No	No	Graph	No
MTS8	Graph	Graph	No	No	Graph	No
MTS9	Yes	Yes	Yes	Yes	Yes	Yes
MTS10	Yes	Yes	Yes	Yes	Yes	Yes
MTS11	Yes	Yes	Yes	Yes	Yes	Yes
MTS12	Yes	Yes	Yes	Yes	Yes	Yes
MTS13	No	No	No	No	No	No
MTS14	No	No	No	No	No	No
MTS15	No	No	No	No	No	No
MTS16	No	No	No	No	No	No
MTS17	No	Graph	No	No	No	No
MTS18	No	Yes	Graph	No	No	Yes
MTS19	No	Yes	Yes	Yes	Yes	Yes
MTS20	No	Yes	Yes	No	No	No
MTS21	No	Yes	Yes	Yes	Yes	Yes
MTS22	No	Yes	Graph	No	Yes	Yes
MTS23	Graph	Yes	Yes	Yes	Yes	Yes

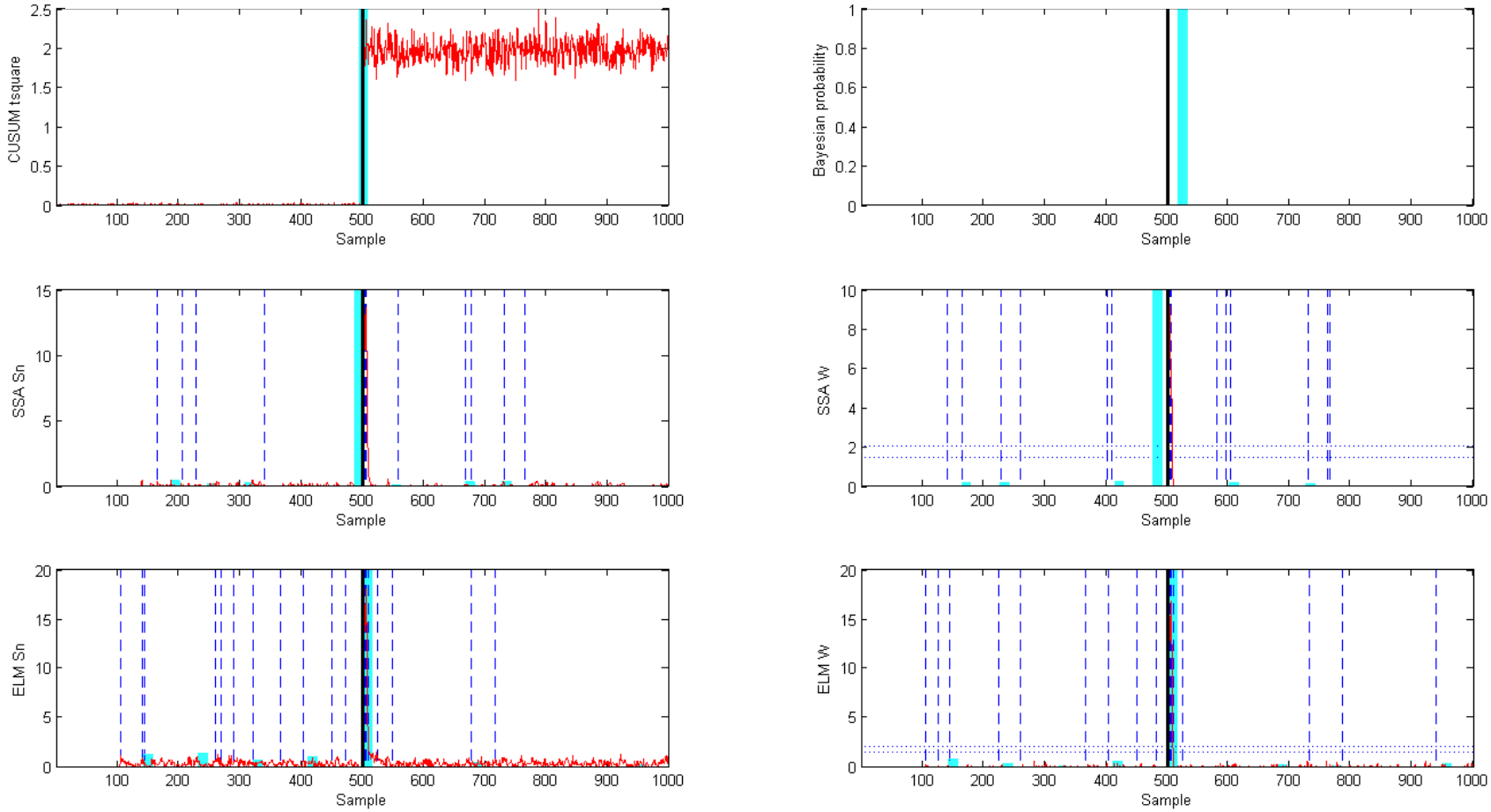

Figure 26: **Simple multivariate time series data: change point detection of fault condition MTS2 (a mean shift in variable x_1 from 0 to 1) at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

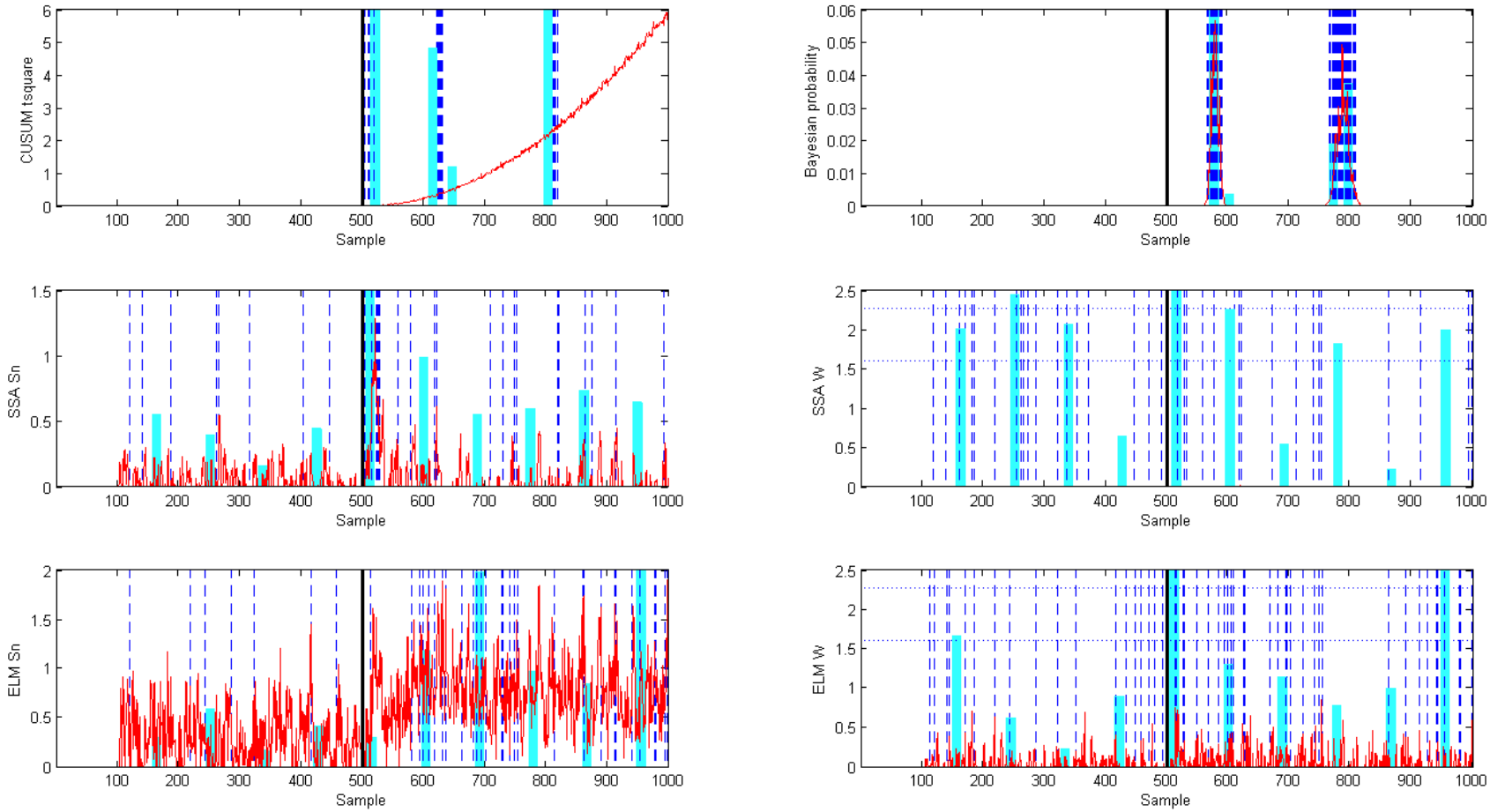

Figure 27: **Simple multivariate time series data: change point detection of fault condition MTS6 (a mean shift in x_1 at a rate of 0.01 per sample) at a confidence level of 0.99**

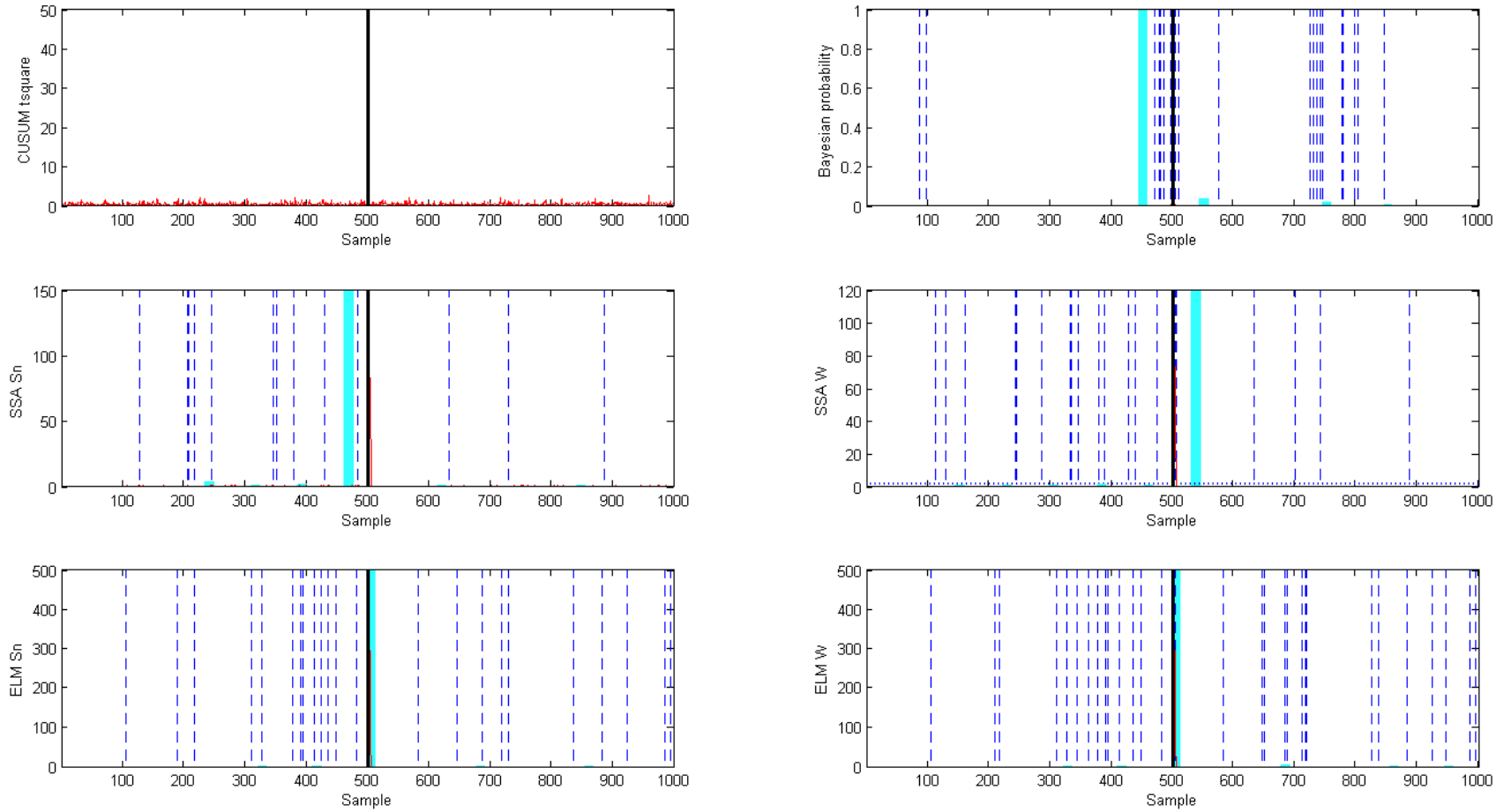

Figure 28: **Simple multivariate time series data: change point detection of fault condition MTS10 (a spike in x_1 having a duration of 10 samples) at a confidence level of 0.99**

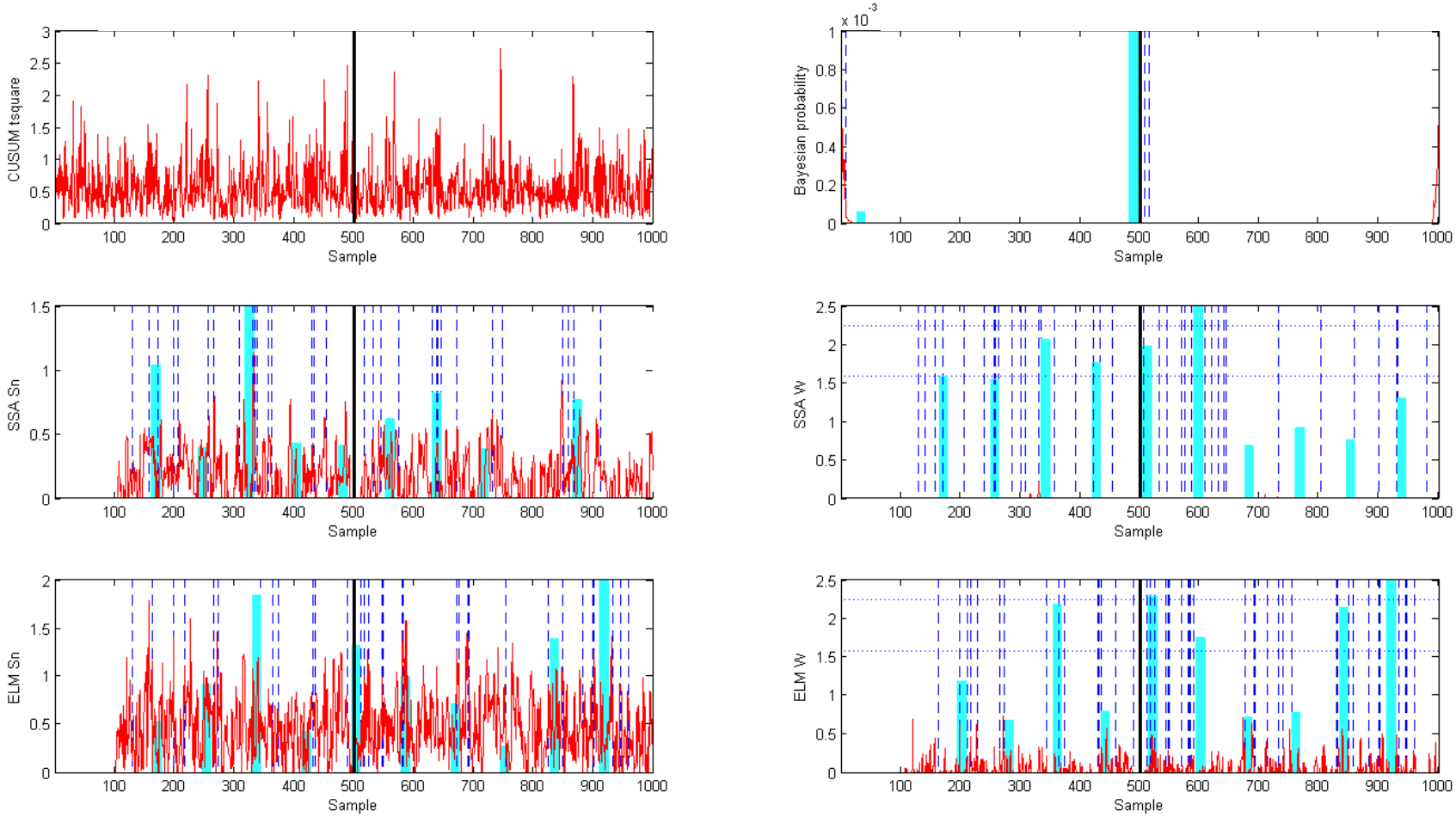

Figure 29: **Simple multivariate time series data: change point detection of fault condition MTS14 (a standard deviation shift in x_1 from 0.1 to 1) at a confidence level of 0.99**

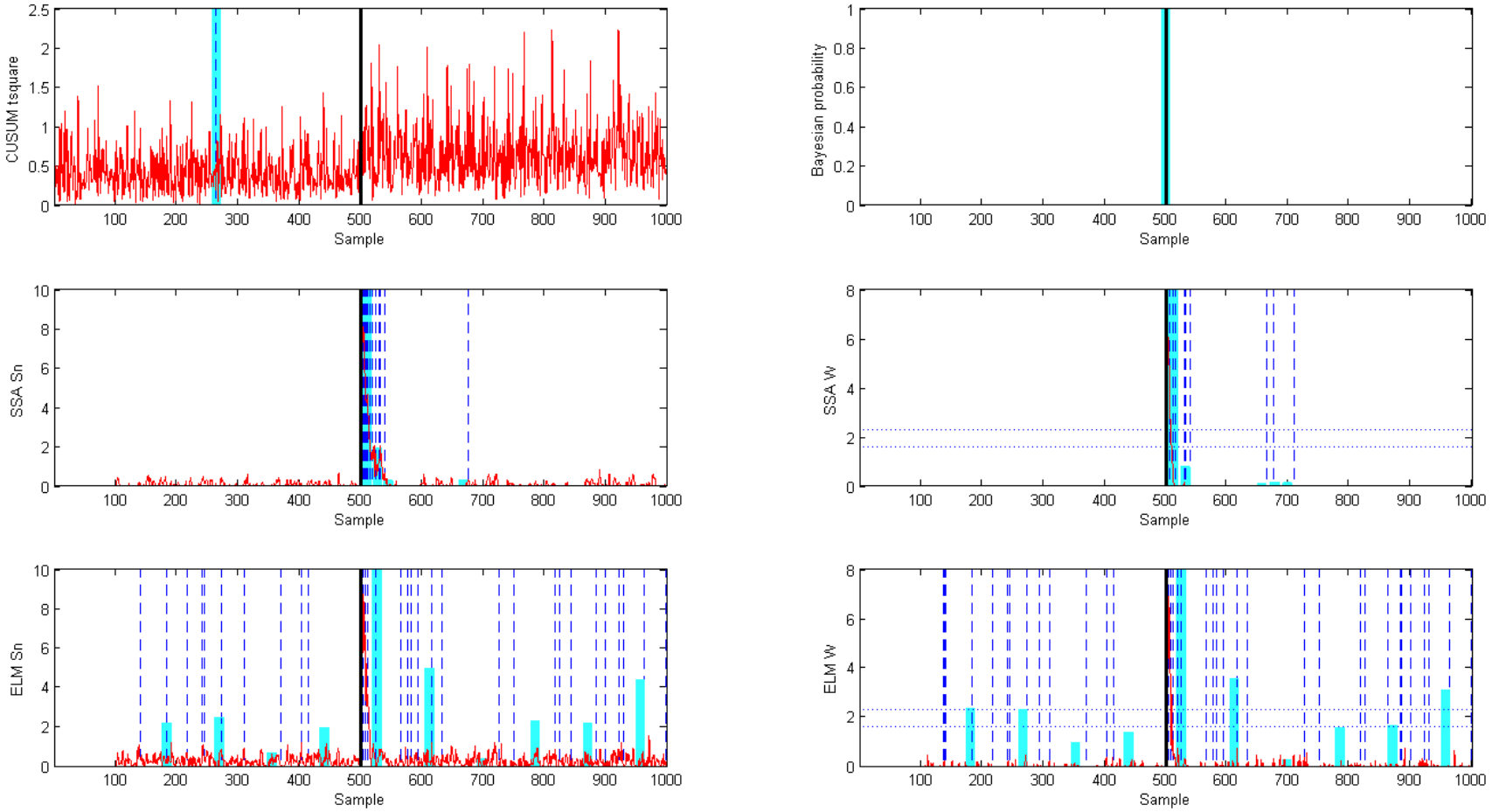

Figure 30: **Simple multivariate time series data: change point detection of fault condition MTS19 (a covariance shift) at a confidence level of 0.99**

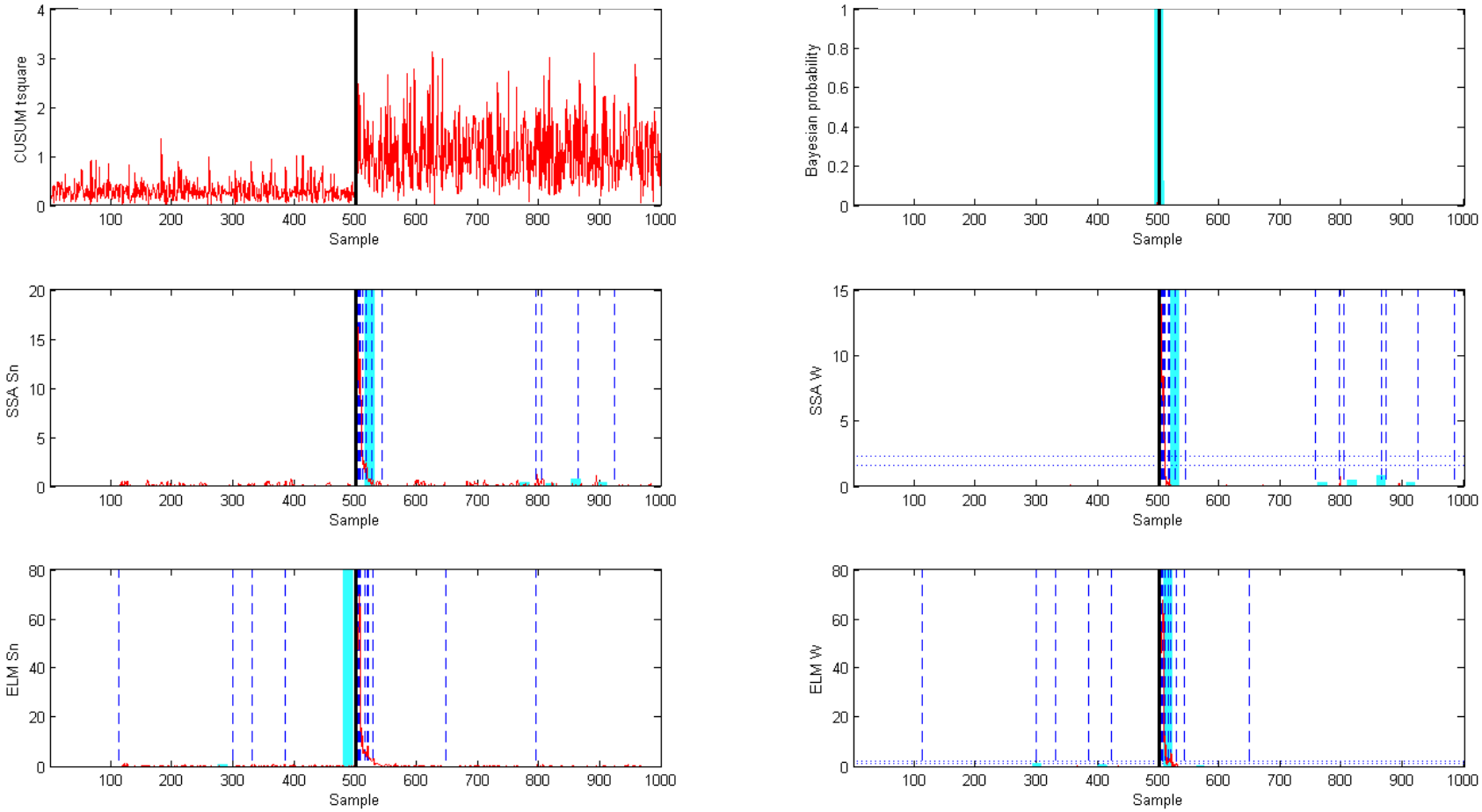

Figure 31: **Simple multivariate time series data: change point detection of fault condition MTS23 (a change in the function type of x_3 to a sine wave with an amplitude of 5) at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

4.3.2 Simple multivariate process

For the simple multivariate process 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 501, indicated by the black vertical line on the change point detection evaluation figures, and the evaluation criteria determined over the following 500 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. Where required, a confidence limit threshold of 99% was used.

As stated previously, inspection of the simple multivariate process data (Figure 11) indicated it to not be independent and identically distributed while having a normal distribution. These data characteristics violate the required assumptions of both the nearest-neighbours CUSUM and Bayesian change point detection techniques, resulting in only the SSA and ELM SSA change point detection techniques being suitable for assessing this particular data set. However, as stated earlier, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Also, the magnitudes of the process changes resulting in the fault conditions are very small and it is expected that all the available techniques will struggle to reliably detect the change points. For the sake of completeness all the available algorithms will therefore be assessed. This will not only allow all appropriate algorithms to be evaluated but also help to gain insight into the robustness of the change point detection techniques whose data assumptions have not been met.

From the analysis of the NOC data (Table 6), false events were only detected for the Bayesian probability statistic. These were, however, easily rejected following visual inspection due to their small magnitude.

Table 6: **Simple multivariate process: NOC results**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
SMP0	No events	Small values	No events	No events	No events	No events

For the mean shift fault conditions, SMP1-SMP5, both the nearest-neighbours CUSUM statistic and the Bayesian probability was only able to detect the change points associated with the larger mean shift fault conditions (Table 7). For the nearest-neighbours CUSUM statistic, the nearest-neighbours CUSUM T^2 was able to visually show the detected change point (Figure 32). The SSA and ELM SSA statistics were unable to detect any of the mean shift fault conditions.

For the parameter change fault conditions, SMP6-SMP8, only the Bayesian probability technique was able to detect the change points associated with the larger parameter change fault conditions (Table 7).

Stellenbosch University http://scholar.sun.ac.za

Neither the nearest-neighbours CUSUM statistic, nor the SSA or ELM SSA statistics were able to detect any of the parameter change fault conditions (Figure 33).

As expected, due to the small magnitudes of the process changes resulting in the fault conditions, most of the available techniques struggled to reliably detect the change points in the data.

Contrary to expectations, it can be concluded that for the simple multivariate process data, not being independent and identically distributed with a normal distribution, the Bayesian probability change point detection technique outperformed all the other techniques tested, correctly detecting 62.5% of the change points. Although the data characteristics of the data set violated the assumptions of the Bayesian probability change point detection technique, the technique proved to be exceptionally robust in this regard as was still able to detect the majority of the change points. The poor performance by the nearest-neighbours CUSUM statistic, only correctly detecting 37.5% of the change points, was as expected due to the multivariate nature of the data set and the complexity of the fault conditions introduced. Contrary to expectation, both the ELM SSA and SSA change point detection techniques failed to correctly detect any of the change points. It is speculated that the small magnitude of the changes introduced as fault conditions, combined with the default parameter selection for the techniques contributed to their poor performance.

Table 7: **Simple multivariate process: fault condition results (Yes = correct change point clearly identified; No = correct change point not identified)**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
SMP1	No	No	No	No	No	No
SMP2	No	No	No	No	No	No
SMP3	Yes	Yes	No	No	No	No
SMP4	Yes	Yes	No	No	No	No
SMP5	Yes	Yes	No	No	No	No
SMP6	No	No	No	No	No	No
SMP7	No	Yes	No	No	No	No
SMP8	No	Yes	No	No	No	No

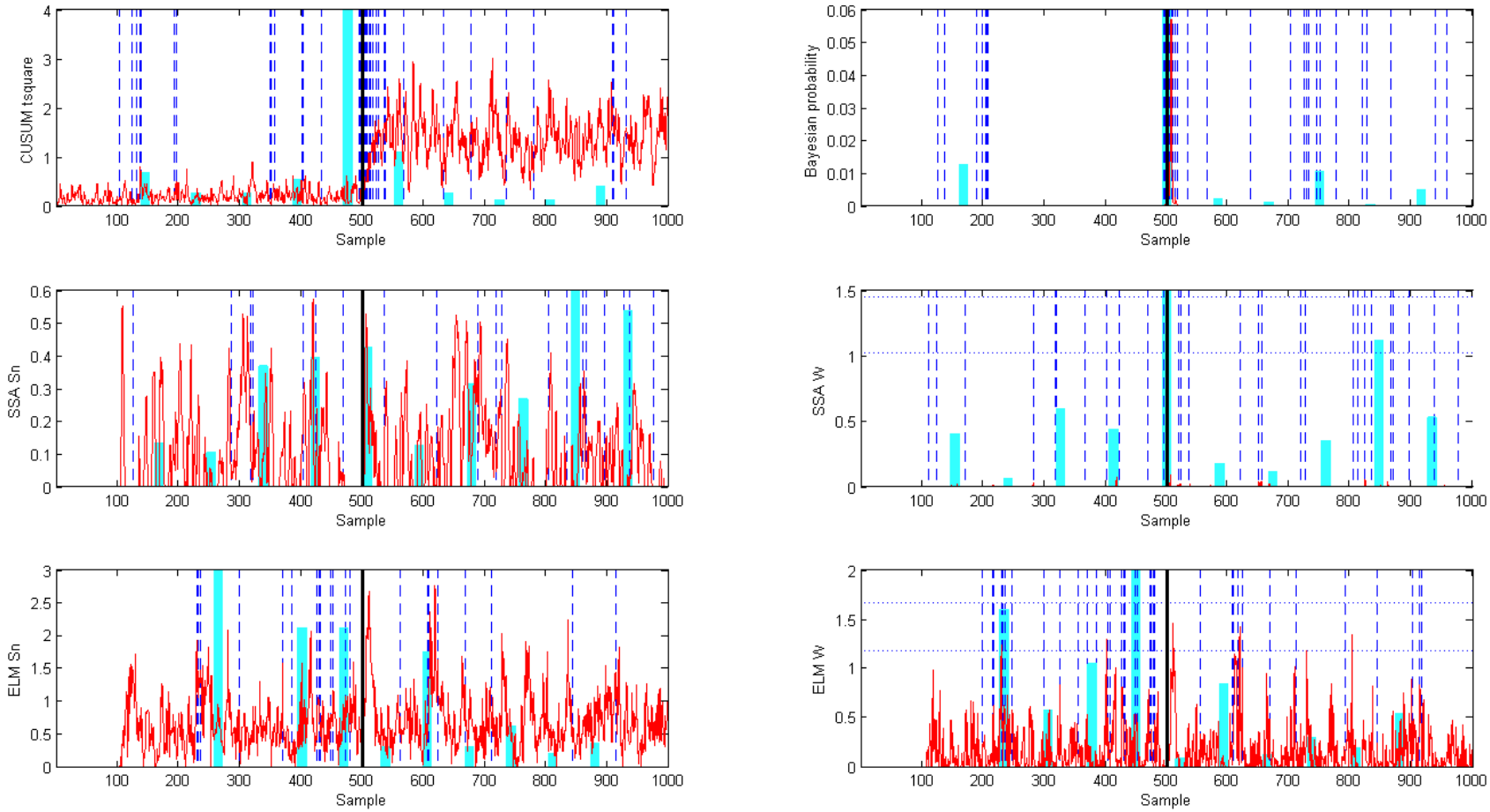

Figure 32: **Simple multivariate process: change point detection of fault condition SMP5 (a step change in the mean of $e_2(1)$ from 0 to 3) at a confidence level of 0.99**

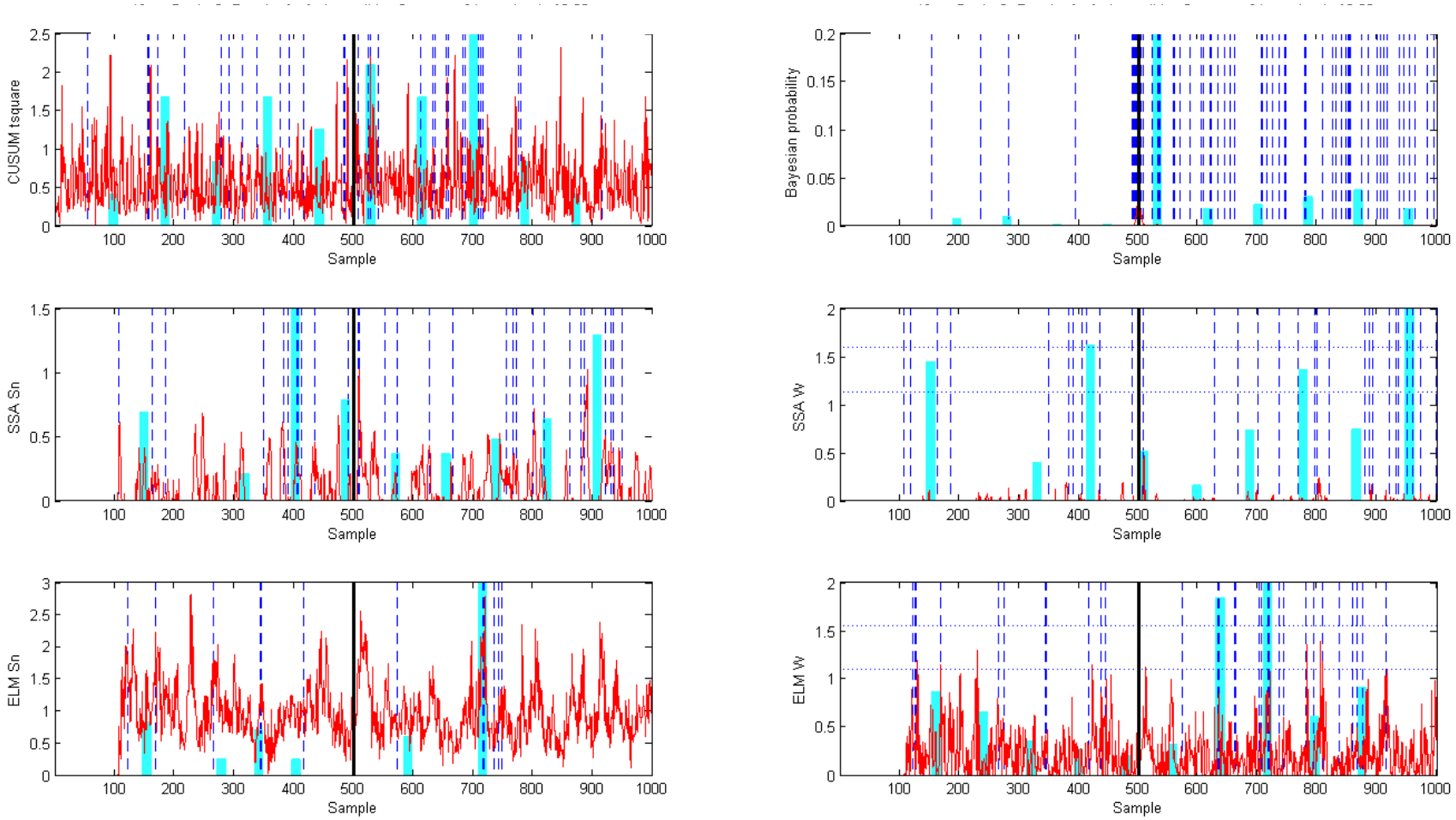

Figure 33: **Simple multivariate process: change point detection of fault condition SMP8 (a step change in the parameters of D_{SMP}(2,1) from 3 to 1) at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

4.3.3 Tennessee Eastman process

For the Tennessee Eastman process 1800 data points at a sampling rate of one sample per three minutes were generated for each of the variables with the fault condition introduced at data point 901, indicated by the black vertical line on the change point detection evaluation figures, and the evaluation criteria determined over the following 900 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. Where required, a confidence limit threshold of 99% was used.

As stated previously, inspection of the Tennessee Eastman process data (Figure 13) indicated it to have a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, with many of the variables having significant autocorrelation with similar profiles. This similarity in the autocorrelation profiles is indicative of the presence of high levels of collinearity (linear correlation structure between the variables). Based on these data characteristics, only the SSA-based change point detection technique would be suitable for assessing this particular data set, with the required assumptions of all the other performance metrics being violated. However, as stated earlier, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Considering that this is quite an extensive data set it is however expected that some of the fault conditions will relate to data that is independent and identically distributed while having a normal distribution, resulting in the other algorithms also being valid some of the time. Therefore, all the available algorithms will be assessed. This will not only allow all appropriate algorithms to be evaluated but also help to gain insight into the robustness of the change point detection techniques whose data assumptions have not been met.

From the analysis of the NOC data (Table 8), false events were only detected for the nearest-neighbours CUSUM statistic. These were, however, easily rejected following visual inspection.

Table 8: **Tennessee Eastman process: NOC results**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
TEP0	False events	No events	No events	No events	No events	No events

It should be noted at this point that for the fault detection technique evaluation of the Tennessee Eastman process it was found that not all the variables responsible for the fault conditions or the variables closely related to the fault conditions were being monitored (Chen and McAvoy, 1998), the control system controlling the Tennessee Eastman process was too quick in correcting some of the induced fault condition, the effect of the fault condition on the process was too slow and the magnitude of some of the fault condition was too small to result in a measurable process performance degradation. These factors could also play a role in the effectiveness of the change point detection techniques being evaluated.

Stellenbosch University http://scholar.sun.ac.za

From the change point detection results (Table 9) the simulated fault conditions can be divided into 3 distinct groups:

a) Fault conditions for which the change points were easily identified by at least one statistic of each of the change point techniques, either visually or from confirmed peaks (Figure 34): fault conditions TEP1, TEP6, TEP4, TEP7, TEP11, TEP12, TEP13, TEP17, TEP22 and TEP23 (showing good overlap with the fault detection results).
b) Fault conditions for which the change points were easily identified by at least one statistic from only two of the change point techniques, either visually or from confirmed peaks: fault conditions TEP2, TEP5, TEP8, TEP10, TEP14, TEP16, TEP18, TEP20, TEP21 and TEP24 (also showing some overlap with the fault detection results).
c) Fault conditions for which the change points were identified by at least one statistic from only one of the change point techniques or not at all (Figure 37), either visually or from confirmed peaks: fault conditions TEP3, TEP9, TEP15 and TEP19 (also showing some overlap with the fault detection results).

The fault conditions in group (a) consist mainly of step changes and “other” disturbances, including two random variation fault conditions. For the change point detection evaluation this group is considerably larger than for the fault detection evaluation, even including a fault condition that was not detectable at all by any of the fault detection techniques (fault conditions TEP12). Unlike for the fault detection evaluation, this group therefore not only included fault conditions being monitored by variables responsible for the fault condition or variables closely related to the fault condition, but also fault conditions whose effect on the process is not that obvious considering the variables being monitored.

The fault conditions in group (b) again consist mainly of step changes and “other” disturbances, including two random variation fault conditions. Unlike for group (a) fault conditions, the size of group (b) for the change point detection evaluation is similar in size that for the fault detection evaluation, however, again including some fault conditions that was not detectable at all by any of the fault detection techniques (fault conditions TEP14 and TEP21). All of the fault conditions in group (b) were detectable by the nearest-neighbours CUSUM change point detection technique, with the Bayesian (Figure 35), SSA (Figure 36) and ELM SSA (Figure 36) change point detection techniques each only being able to detect about half of the fault conditions (Bayesian: fault conditions TEP2, TEP10, TEP14, TEP16, TEP21 and TEP24. SSA: fault conditions TEP5, TEP8, TEP16, TEP18 and TEP20. ELM SSA: fault conditions TEP2, TEP5, TEP8, TEP18, TEP20, TEP21 and TEP24). Following visual inspection of the results it was noted that there was a considerable delay in detecting the fault conditions TEP19, TEP20 and TEP21. This can be ascribed to the fact that the variables inducing the fault conditions is not being directly monitored and the effect on the process only becomes evident over time considering the variables being monitored.

The fault conditions in group (c) consist of 2 step changes, a random variation change and an “other” disturbance. For this group, only fault condition TEP19 was detectable and only using the SSA change point detection technique. As with the Tennessee Eastman process fault detection technique evaluation,

Stellenbosch University http://scholar.sun.ac.za

this can be ascribed to the fact that variables responsible for the fault conditions or variables closely related to the fault conditions were not being monitored (Chen and McAvoy, 1998), the control system controlling the Tennessee Eastman process was too quick in correcting the induced fault condition, the effect of the fault condition on the process was too slow or the magnitude of the fault condition was too small to result in a measurable process performance degradation over the evaluation period.

Contrary to expectations, it can be concluded that for the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, that very similar performance was achieved by all the change point detection techniques evaluated when excluding cases where visual inspection was required. Although the data characteristics of the data set violated the assumptions of all but the SSA change point detection technique, all the techniques proved to be exceptionally robust in this regard as they were all able to detect a reasonable number of the change points. Whereas the nearest-neighbours CUSUM change point detection technique required the most visual inspection of results for confirmation (having been able to correctly detect 79.2% of the change points), running the Bayesian and SSA change point detection techniques in parallel will deliver the best results regarding confirmed peaks (being able to correctly detect 87.5% of the change points). Having a non-linear structure and monitoring residuals, the ELM SSA change point detection technique performed fairly well considering the high degree of autocorrelation, and consequently high levels of collinearity, in the data, with the technique being able to correctly detect 70.8% of the change points (slightly better when compared to 66.6% for the Bayesian and SSA change point detection techniques). Although the change point detection techniques performed slightly better than the fault detection techniques, it was again found that the Tennessee Eastman process poses a very complex problem. As mentioned for the fault detection technique evaluation, it is again suggested that an approach to potentially reducing the complexity inherent to monitoring the Tennessee Eastman process as a whole lies with the idea of simplifying the problem through the use of process causality maps.

Stellenbosch University http://scholar.sun.ac.za

Table 9: **Tennessee Eastman process: fault condition results (Yes = correct change point clearly identified; Graph = correct change point visually inferred; No = correct change point not identified)**

Fault condition	CUSUM T^2	Bayesian probability	SSA S_n	SSA W	ELM S_n	ELM W
TEP1	Yes	Yes	Yes	Yes	Yes	Yes
TEP2	Yes	Yes	No	No	Yes	Yes
TEP3	No	No	No	No	No	No
TEP4	Yes	Yes	Yes	Yes	Yes	Yes
TEP5	No	No	Yes	Yes	Yes	Yes
TEP6	Yes	Yes	Yes	Yes	Yes	Yes
TEP7	Yes	Yes	Yes	Yes	Yes	Yes
TEP8	Yes	No	Graph	Yes	Yes	Yes
TEP9	No	No	No	No	No	No
TEP10	Yes	Yes	No	No	No	No
TEP11	Yes	Yes	Graph	No	Graph	Yes
TEP12	Graph	Yes	Graph	Yes	Graph	Yes
TEP13	Graph	Yes	Yes	Yes	Graph	Yes
TEP14	Graph	Yes	No	No	No	No
TEP15	No	No	No	No	No	No
TEP16	Graph	Yes	No	Yes	No	No
TEP17	Yes	Yes	Yes	Yes	Yes	Yes
TEP18	Yes	No	Yes	Yes	Yes	Yes
TEP19	No	No	No	Yes	No	No
TEP20	Yes	No	Yes	Yes	Yes	Yes
TEP21	Graph	Yes	No	No	Yes	Yes
TEP22	Yes	Yes	Yes	Yes	Yes	Yes
TEP23	Yes	Yes	Yes	Yes	Yes	Yes
TEP24	Yes	Yes	No	No	Yes	Yes

Stellenbosch University http://scholar.sun.ac.za

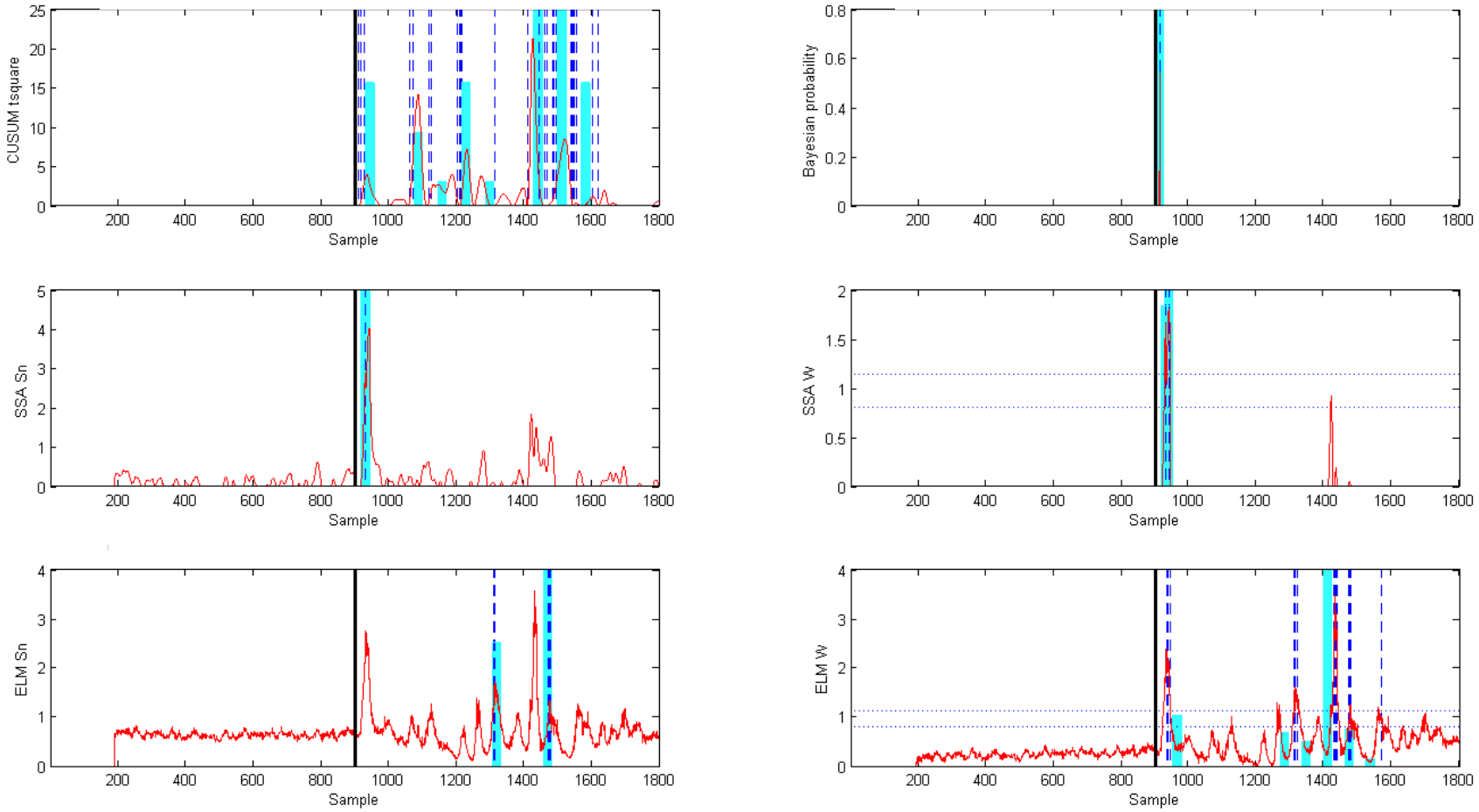

Figure 34: **Tennessee Eastman process: change point detection of fault condition TEP13 (a slow drift in the reaction kinetics) at a confidence level of 0.99**

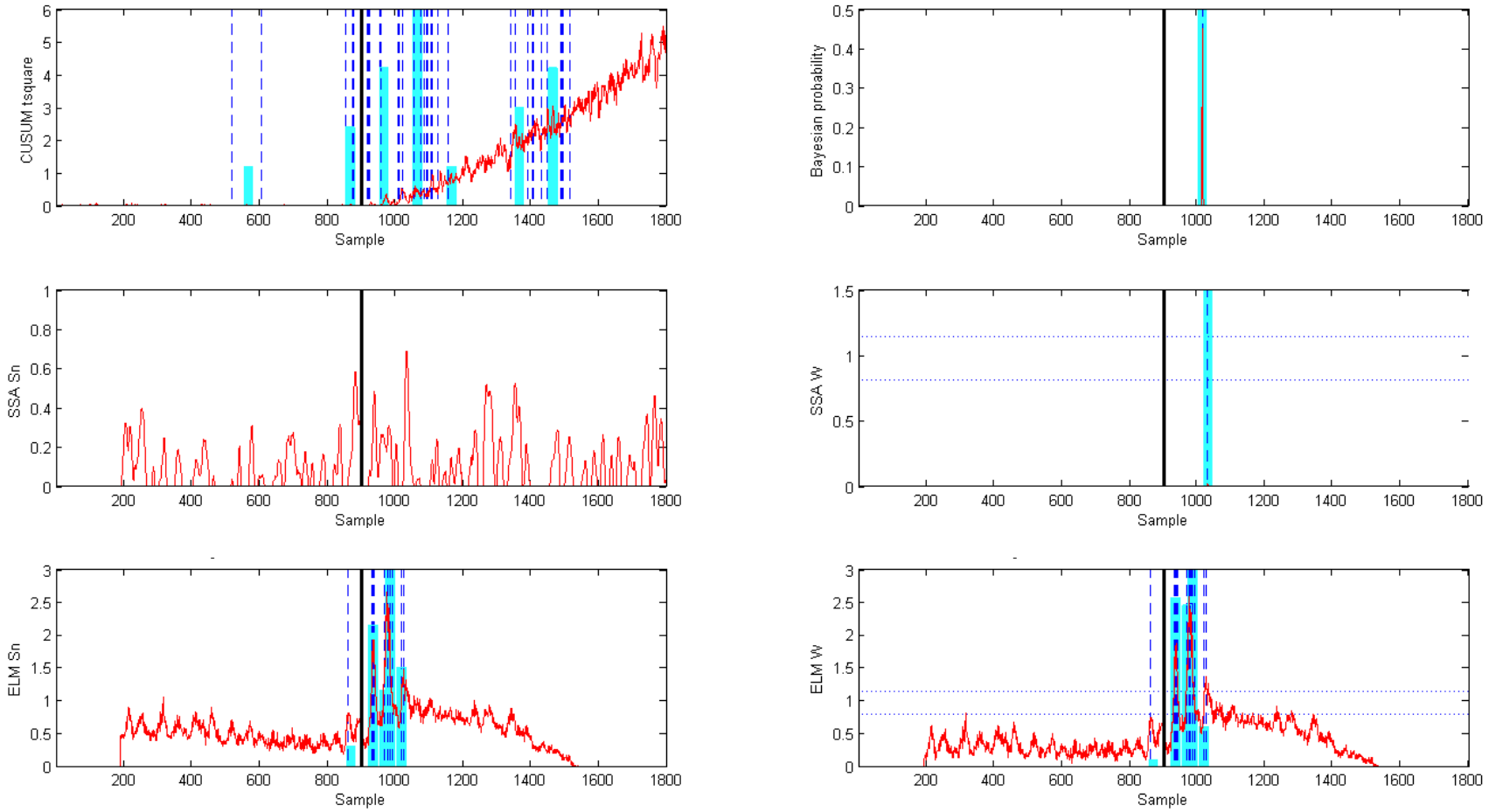

Figure 35: **Tennessee Eastman process: change point detection of fault condition TEP21 (a decrease of 15% in the production rate) at a confidence level of 0.99**

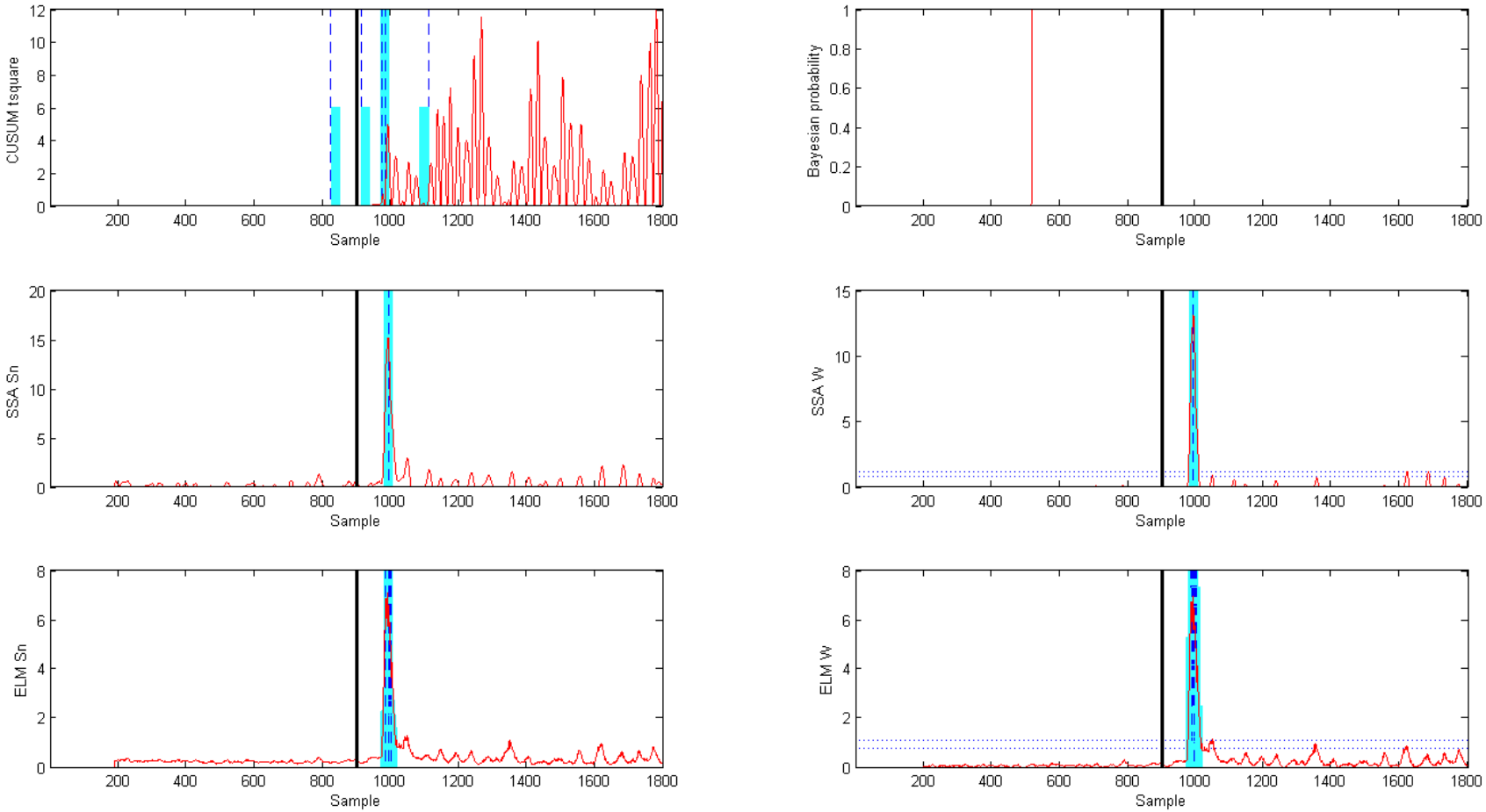

Figure 36: **Tennessee Eastman process: change point detection of fault condition TEP20 (an unknown disturbance) at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

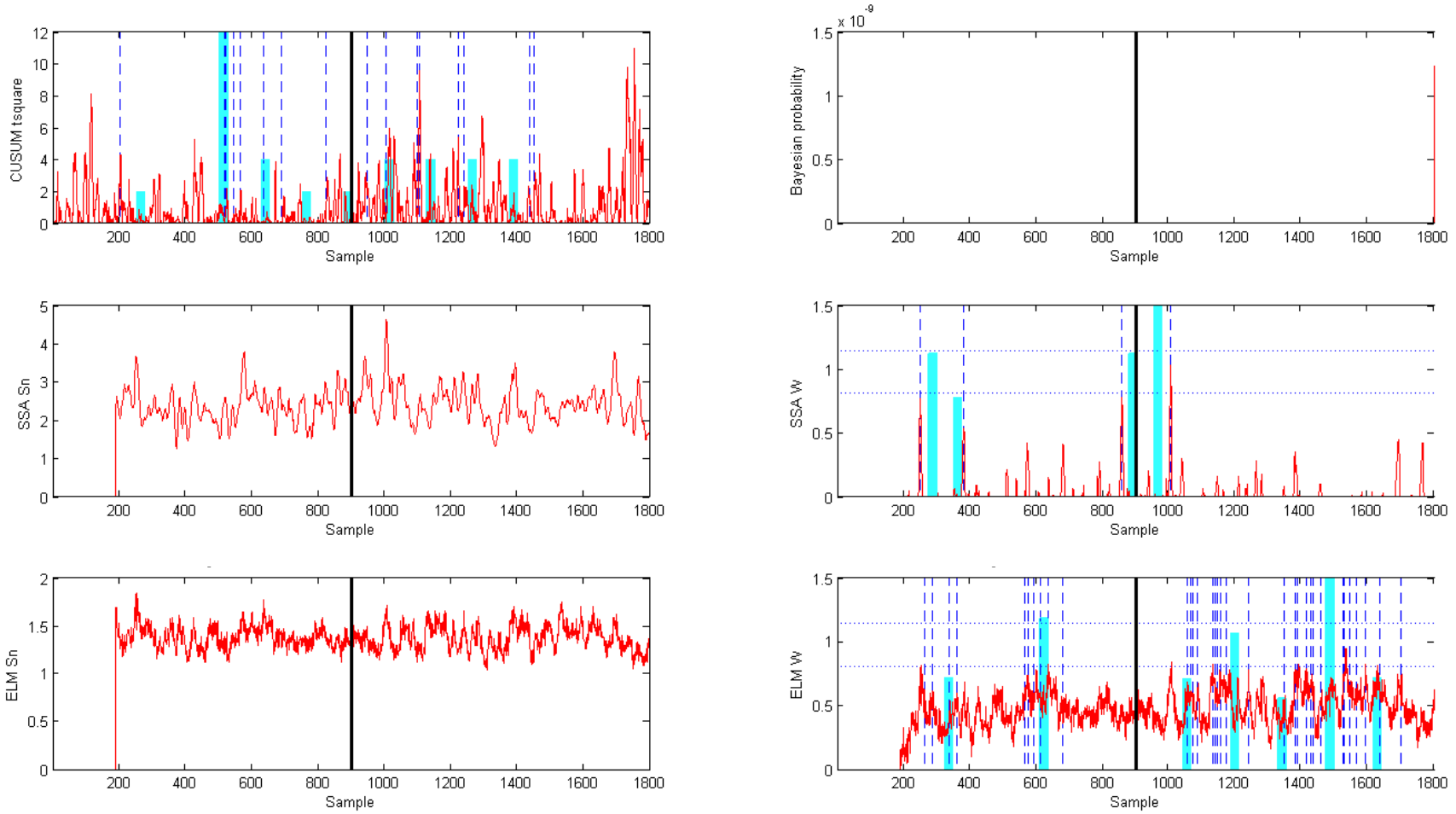

Figure 37: **Tennessee Eastman process: change point detection of fault condition TEP9 (a random variation in the D_{TE} feed temperature) at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

4.3.4 Summary

From the change point detection technique evaluation it is evident that there is no single change point detection technique that is effective in identifying all potential fault conditions. It was found that all the techniques performed very similarly, however, the nearest-neighbours CUSUM change point detection technique required the most validation, in the form of visual inspection, to confirm results. The nearest-neighbours CUSUM change point detection technique is also actually a univariate technique, having only multivariate results representation. Due to the nature of the data being analysed, different techniques may be required considering the assumptions the techniques are based upon:

- For the simple multivariate time series data, being independent and identically distributed with a normal distribution, the Bayesian change point detection technique performed the best of the techniques tested.
- For the simple multivariate process data, not being independent and identically distributed with a normal distribution, the Bayesian probability change point detection technique again outperformed all the other techniques tested.
- Furthermore, although the data characteristics of the data set violated the assumptions of the Bayesian probability change point detection technique, the technique proved to be exceptionally robust in this regard as was still able to detect the majority of the change points.
- For the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, very similar performance was achieved by all the change point detection techniques evaluated.
- As before, although the data characteristics of the data set violated the assumptions of all but the SSA change point detection technique, all the techniques proved to be exceptionally robust in this regard as they were all able to detect a reasonable number of the change points.

From this it is evident that the Bayesian change point detection technique was found to be especially robust, even when the assumptions the technique is based upon are violated, making it the algorithm of choice for change point detection. The effectiveness of the algorithm could potentially be improved by pre-processing the data to remove autocorrelation when present. However, as with the fault detection techniques, it is suggested that multiple change point analysis techniques be run in parallel and their results be interpreted in conjunction with expert process knowledge.

Following the success of the ELM PCA algorithm for fault detection, it was proposed to implement the ELM algorithm as a residual generation stage, with the aim of removing the non-linear and dynamic characteristics from the data, prior to SSA change point detection. This permitted the effective application of the linear SSA change point detection algorithm to non-linear data. The key characteristic of the ELM algorithm that made this possible is its extremely fast learning speed. Due to the way in which the SSA change point detection algorithm requires the continual updating of its base and test data matrices, the use of more slowly trained conventional feedforward neural networks is not possible due, especially when considering on-line applications. It was found that the performance of the proposed ELM SSA change point detection technique was slightly better than the SSA change point detection technique on which it is

Stellenbosch University http://scholar.sun.ac.za

based. In this regard, the ELM SSA change point detection technique was more effective at detecting more subtle changes in the data, such as a slowly shifting mean in the data (drift), when compared to the SSA change point detection technique

When comparing the change point detection results to the statistical data-based fault detection results, it is evident that the change point detection techniques are at least as capable at detecting fault conditions as the statistical data-based fault detection techniques are, in some instances even being able to detect fault conditions that the statistical data-based fault detection techniques could not. From this is concluded that the change point detection techniques are a viable alternative to the statistical data-based fault detection techniques. The advantage of the change point detection techniques over the statistical data-based fault detection techniques lies with the fact that they do not depend on fixed NOC data. Change point detection aims to determine whether or not a process in its current state is exhibiting a change in behaviour compared to the behaviour it exhibited immediately preceding its current state, irrespective of whether it was in a state of normal process behaviour or not. This continuous updating of the “normal” operating condition data allows the change point detection techniques to be used to not only identify changes in process behaviour when changing from normal to abnormal process operation, but also changes in process behaviour while being in a normal or abnormal process state. From an application perspective this has the benefit that no NOC data has to be defined upfront for the change point detection techniques to be used. The results from applying change point detection can be used to partition the data into different process states that can subsequently be monitored using statistical data-based fault detection techniques.

A potential enhancement, especially applicable to the nearest-neighbours CUSUM change point detection technique is to first apply some of the fault detection techniques to the data (considering all data as NOC data). The fault detection technique will effectively be used for dimensionality reduction and feature extraction, allowing the change point detection technique the detect changes in the fault detection space. As such, the change point detection techniques can potentially be used to enhance the fault detection thresholds, which are based on NOC data, something the change point detection techniques are less concerned about. Also, following the detection of change points, fault detection techniques can be used to validate the findings. As with the fault detection techniques, it is again suggested that for complex processes having many variables requiring monitoring, the complexity of the problem be reduced through the development of process causality maps.

Stellenbosch University http://scholar.sun.ac.za

5 Variable importance analysis

Whereas SPC contribution plots are the predominant approach used in identifying the variable(s) responsible for process fault conditions identified through statistical data-based fault detection there are situations when contribution plots are not suitable, possibly due to the underlying fault detection model being inaccurate, or not reliable enough for use. Also, an alternative to SPC contribution plots is needed for analysing potentially interesting process events detected by change point detection. In this context, variable importance analysis complements SPC variable contribution plots in that it allows for the analysis of important variables where the application of fault detection models may be unsuitable or unreliable.

In many applications it is often more important to identify the relative importance of input variables, rather than only predicting the response by means of some "black-box" model. Variable importance analysis makes use of variable importance measures to compare candidate input variables with respect to their influence on predicting the response or even causal effect. In doing so, variable importance measures give a ranked indication of the relative significance of input variables to a classification or regression problem. Three groups of methods are available to address problems requiring variable importance analysis (Zhou et al., 2006):

- Model-based methods – expensive and require accurate analytical process models.
- Expert system-based methods – flexible and fast but difficulties can arise when there is a lack of sufficient process knowledge.
- Data-driven methods – dependent on large amounts of process data and their usefulness is highly dependent on the quality of the data.

For this study, data-driven methods that evaluate problems described as supervised classification problems will be used. For supervised learning, input data is mapped to output data using a model. The model can be either a classification type model (for categorical output data) or a regression type model (for continuous output data). For classification, each category of the output data represents a different process condition or state experienced.

Important considerations for variable importance analysis include variable selection and variable correlations. Chiang et al. (2004) showed that choosing key variables before applying classification techniques leads to improved classification and more accurate fault diagnosis. When all variables in a data set are used, irrelevant variables may be included in the analysis, negatively affecting the prediction by camouflaging the significant features, thus leading to poor classification and ultimately poor fault diagnosis. For data sets containing correlated independent variables it is often more difficult for the analysis technique to correctly predict the most important independent variable. In these cases the most important variable will be ranked amongst the variables with the highest importance value, however, a covariate correlated to the correct variable may be predicted as the most important variable. It has been shown by Archer & Kimes (2008), that variable importance analysis using random forests overcomes this problem.

Stellenbosch University http://scholar.sun.ac.za

Useful metrics for evaluating different variable importance models are (Sathyanarayanamurthy and Chinnam, 2009):

- Accuracy – measured in terms of mean squared error (MSE).
- Robustness – measured in terms of the capability of achieving good accuracy for different problem types and sample sizes.
- Efficiency – measured in terms of the computational effort required for constructing the model and for predicting the response for a new data set.
- Conceptual simplicity – measured in terms of the ease of implementation.

Conventionally, variable importance analysis can be considered to be analogous to sensitivity analysis. For sensitivity analysis, regression techniques such as step-wise regression are used to identify the model input variables to which a model output variable is most sensitive. Such analyses are usually performed for data representing a specific state or class and can be considered within-class analyses. For this study the requirement is for the analysis of variable importance between different states or classes, hence the need for between-class variable importance analysis with the focus on discrimination and/or classification techniques. Similar to step-wise regression, combining a classification model with a leave-one-out algorithm can be used for between-class variable importance analysis. To this end the techniques selected for evaluation are based on linear discriminant analysis (LDA), classification trees (CT) with bagging (trees-with-bagging), random forests (RF), classification extreme learning machine (ELM) with bagging (ELM-with-bagging) and canonical variance analysis (CVA) biplots. Similar to PCA for statistical data-based fault detection, linear discriminant analysis, being a relatively basic linear technique for discriminating between different data classes, forms the basis of the variable importance analysis evaluation. Classification trees, being a non-linear distribution free technique, can be considered an improvement over LDA, especially when considering the classification of non-linear data. Ensembles of classification trees, such as random forests, have been shown to significantly improve the classification accuracy of classification trees (Breiman, 2001) through the generation and aggregation of multiple classification trees. Subsequently, extreme learning machine for classification, having been shown to be good at generalisation and having extremely fast learning speeds with very few parameters that require setting (Huang et al., 2006), are considered as a viable alternative to the aforementioned classification techniques. Lastly, CVA biplots, a powerful data visualisation technique, have successfully been used to determine discriminating process conditions between different process states (Gardner et al., 2005).

As with the change point detection techniques, the variable importance analysis techniques under consideration can be divided into procedural and interactive techniques. Whereas procedural techniques require only the specification of parameters and input data in order to arrive at usable results, interactive techniques require some inspection and possible adjustment of the results before any conclusions can be drawn.

This chapter deals with some of the theory behind variable importance analysis and its application to process performance monitoring. The selected techniques are not an exhaustive list of available techniques, but have been chosen to be illustrative of what is available.

Stellenbosch University http://scholar.sun.ac.za

5.1 Procedural variable importance

5.1.1 Linear Discriminant Analysis

LDA is a well-known classical statistical method used to find linear combinations of features which separate multiple classes the best. LDA is a version of Fisher discriminant analysis (FDA) under the assumptions of normally distributed classes or equal class covariance. FDA, a linear dimensionality reduction technique, is optimal in terms of maximising the separation between multiple classes and has been shown to be the optimal linear technique for fault diagnosis (Chiang et al., 2001).

FDA finds linear or discriminant functions that yield a new set of transformed values, providing a more accurate discrimination than the original variables. The function is used to separate the data by projecting it onto transformed axes that maximise the ratio of variance between classes to variance within classes. Because these functions try to separate data, there is a strong link between discrimination and classification. First, the total scatter matrix, SC, is derived by stacking the training data for all classes into a $n \times p$ matrix X and representing the j th row of X with the column vector x_j (Chiang et al., 2004):

$$SC_t = \sum_{j=1}^{n} \left(x_j - x_{mean}\right)\left(x_j - x_{mean}\right)^T \tag{5-1}$$

where x_{mean} is the total mean vector whose elements correspond to the means of the columns of X. Next, define the set of vectors x_j which belong to the class k as X_k, then, for class k, the within-scatter is defined as:

$$SC_k = \sum_{x_j \in X_k}^{n} \left(x_j - x_{k,mean}\right)\left(x_j - x_{k,mean}\right)^T \tag{5-2}$$

where $x_{k,mean}$ is the mean vector for class k. With δ as the number of classes, the within-class-scatter matrix:

$$SC_w = \sum_{k=1}^{\delta} SC_k \tag{5-3}$$

and the between-class-scatter matrix:

$$SC_b = \sum_{k=1}^{\delta} n_k \left(x_{k,mean} - x_{mean}\right)\left(x_{k,mean} - x_{mean}\right)^T \tag{5-4}$$

Stellenbosch University http://scholar.sun.ac.za

can be defined, where n_k is the number of observations in class k. From this, the first FDA vector, l_1, can now be determined as:

$$\max_{l_1} \frac{l_1^T SC_b l_1}{l_1^T SC_w l_1} \tag{5-5}$$

Each of the subsequent FDA vectors are then computed so as to maximise the scatter between classes while minimising the scatter within classes among all axes perpendicular to its preceding FDA vector. These FDA vectors are the eigenvectors, l_k, of the generalised eigenvalue problem:

$$SC_b l_k = \lambda_k SC_w l_k \tag{5-6}$$

where the eigenvalues, λ_k, specify the separation between classes by projection onto l_k. Using discriminant analysis, observations can now be classified according to this reduced FDA space (Hastie et al., 2009).

Once the class prediction is determined a misclassification error can be calculated. The misclassification error is determined by comparing the predicted response to the actual response for each observation. The total number of observations falsely classified, as a ratio of the total number of observations, determines the misclassification error. The misclassification error is used to determine the most important variables. For the variable importance algorithm, the disciminant analysis technique is applied together with a leave-one-out algorithm. The leave-one-out algorithm effectively removes an input variable from the data set and applies the discriminant analysis. This is repeated for all the variables present. For each repetition, a different variable is left out of the analysis. When an independent variable is removed, the corresponding misclassification error contains the effect on the response when the variable is ignored. Therefore, the variable associated with the largest misclassification error is deemed to be the most important variable.

As with dynamic PCA (Ku et al., 1995), for auto-correlated data, incorporating time lags into the variables is an effective way of capturing the dynamics in the data. Incorporating time lags for auto-correlated data in classification problems is seen as a method of decreasing the overlap of the classes with added dimensions (Chiang et al., 2004). Incorporating time lags into the data set is expected to perform well as long as there is enough data to justify the added dimensions. A disadvantage of adding the extra dimensions is an increase in complexity, however, this should be minimal if time lags are only added after key variables have been selected for the analysis.

Whereas discriminant analysis is used in the algorithm to determine between-class variable importance, replacing it with linear regression, within-class variable importance can be calculated. However, linear regression is often misapplied when determining variable importance due to the fundamental

Stellenbosch University http://scholar.sun.ac.za

assumptions of the technique. A key assumption of linear regression is that explanatory variables that take part in the regression are uncorrelated amongst them i.e. there exists no collinearity amongst the explanatory variables. In practice, however, there often exists a remarkable degree of collinearity within a selection of explanatory variables. The effect of high collinearity is an increased variance in the regression input space and a redundancy of some explanatory variables, even though the regression model could still be statistically significant. In engineering terms such a situation could lead to mistaken identification of important variables.

Testing for collinearity can be accomplished as follow:

- Check if the rank of the data is less that the number of explanatory variables.
- Principal component analysis can be applied to the selection of explanatory variables, X. Plotting the eigenvalues of the covariance of X can give a qualitative idea of the potential redundancy amongst explanatory variables in X i.e. a prominent difference in value between the first and tail end of eigenvalues could indicate possible collinearity. A quantitative check can be obtained from the condition index of each principal component. A high condition index, beyond 30 (Aldrich, 2002) may also indicate collinearity amongst input variables.
- Stronger evidence of collinearity can be reached in a combinatorial linear regression amongst all input variables. A permutation of linear regression is performed on each of x_j in X of the remainder x_k in X. That is, $xj = X_k \times b$ where X_k is the set excluding x_j with b a linear vector of coefficients. Any evidence of significant linear relation between any x_j and x_k will indicate collinearity within X. Furthermore, the variance inflation factor (VIF) is also calculated for each x_j. If the VIF is greater than 10, it shows strong evidence of collinearity between the particular variable and the other input variables. Examining the standardised regression coefficient estimates for each combination of regression gives a suggestion as to which variables and to what extent they are related.

5.1.2 Trees-with-Bagging

Trees-with-bagging is based on the method for developing bagging predictors (Breiman, 1996). A bagging predictor constitutes an ensemble of base learners, each developed using a data set randomly sub sampled with replacement from the original. For the trees-with-bagging approach, the base learners are classifications trees, where the aim is to determine between-class variable importance, and regression trees, where the aim is to determine within-class variable importance. By sequentially removing individual variables from the model learning process, it is possible to assess their importance with regards to their contribution to class specific information.

Although many tree-structured classification and regression algorithms are available, Classification and Regression Trees (CART) is used in the trees-with-bagging method. In the general classification problem it is known that each case in a sample belongs to one of a finite number of possible classes, and given a

Stellenbosch University http://scholar.sun.ac.za

set of measurements for a case, it is desired to correctly predict to which class the case belongs (Sutton, 2005).

CART (Breiman et al., 1984) solves this general classification problem by growing a tree structure, portioning the data into mutually exclusive groups (nodes) each as pure or homogenous as possible concerning their response variable. The root node of such a tree contains all the objects which are subsequently divided into nodes by recursive binary splitting as you move down the tree. For each node the binary split is defined by a simple rule based on a single explanatory variable. It should be noted that following a split into two subsets, the resulting subsets do not both subsequently have to be divided using the same variable, allowing non-homogenous responses to be modelled. Also, a classification tree does not have to be symmetric in its pattern of nodes. Questier et al. (2005) summarises the CART algorithm as follow:

1. All objects are assigned to a root node.
2. Each explanatory variable is split at all its possible split points.
3. For each split point, the objects with values higher and lower than the split point for the considered explanatory variable is separated, splitting the parent node into two child nodes.
4. The split point and variable with highest reduction of impurity is selected.
5. Split the parent node into the two child nodes based on the selected split point.
6. Repeat steps 2-5, considering each node as a new parent node and continue until the tree has reached is maximum size.
7. Using cross-validation, prune the tree back to its optimal size.

The measure of impurity used to determine the best variable and split point for each node is the Gini index of diversity (Breiman et al., 1984) and can be defined as:

$$Gini = 1 - \sum_{k=1}^{\delta} \left(\frac{n_k}{n} \right)^2 \qquad (5\text{-}7)$$

where n is the number of objects, δ is the possible classes and n_k is the number of objects from class k present in the node.

Pruning of a developed tree of maximum size is required because such a tree is usually over-fitted. Trees of maximum size have typically fitted every idiosyncrasy, including noise, in the learning data set, all of which are not likely to be found in similar form in future data. Pruning consists of removing branches from the tree resulting in the smallest decrease in accuracy compared to removing other branches. For this purpose a cost-complexity measure is defined for each subtree (Questier et al., 2005):

$$B_{\zeta}(\varphi) = B(\varphi) + \zeta |\varphi| \qquad (5\text{-}8)$$

Stellenbosch University http://scholar.sun.ac.za

where φ is the subtree, $|\varphi|$ is the complexity of φ (number of terminal nodes), ζ is the complexity parameter and $B(\varphi)$ is the overall misclassification rate for classification trees. Each value of ζ will be associated to a unique smallest tree, minimising the cost complexity measure. As ζ is increased from a value of 0, the size of the resulting tree will decrease. Therefore, choosing the best size tree could be used to define the best tree. Cross validation could be used to determine the optimal tree size. For cross validation, the data set is randomly divided into O subsets. The tree growing and pruning procedure is then repeated O times, each time using a different subset as test set and combining the remaining $O-1$ subsets as the training set. The classification error is calculated for each tree size, averaged over all the subsets and then matched with the results for the subtrees of the complete data set using the ζ values. The tree with the lowest cost-complexity measure will be the optimal sized tree.

Tree-structured classification and regression methods are not based on assumptions of normality and user-specified model statements and can be applied to data sets having both large numbers of cases and large numbers of variables, making them extremely resistant to outliers (Steinberg and Colla, 1995). Tree-based models can also always predict the response variable with absolute accuracy for any given training data set. However, such a model will be complex and not capture the trends in the data. When given an entirely new set of data of similar structure, the model will perform poorly – i.e. the model will over-fit the data. When using tree techniques, and attempting to develop tree models, the balance of accuracy and complexity always needs to be considered. Ensemble algorithms and cross validation are employed in an attempt to balance the accuracy and complexity of a model and thus alleviate the problem of over-fitting a model to the given data.

Furthermore, although tree-based models are often seen as easy to interpret, the instability of trees, where very small changes in the learning sample values can result in significant changes in the variables used for the splits, can prevent firm conclusions from being reached when considering problems such as overall variable importance by merely examining the developed tree (Sutton, 2005). This stems from the fact that for highly correlated variables if one is used at an early stage during the model training, there may be little necessity for using the other variables at all. This behaviour should, however, not be taken as evidence that these variables are not strongly related to the model response. Similarly, important interactions can be hard to identify due to correlations between predictor variables. It should also be noted that classification trees produced using the CART algorithm are not guaranteed to be optimal. This is a consequence of the fact that at each stage during the tree growing process, the selected split is based on the one which will immediately reduce the impurity the most – a greedy algorithm – and not based on setting things up for further splitting to be more effective.

Bagging is a general technique for improving model stability and predictive power through the reduction of variance associated with prediction (through averaging for regression and simple voting for classification) and can be applied to tree-based models to increase the accuracy of the resulting predictions. Bagging can be particularly effective when using a generally unstable modelling technique where the correct form

of the model is complicated and unknown, many predictor variables exist (with some being completely unrelated to the response variable) and the sample size is not too big (Sutton, 2005). Bagging a CART classifier can often make it close to being ideal where the misclassification rate will be close to the Bayes rate. This results from the fact that a not too small carefully created classification tree will typically have a relatively small bias and possibly a large variance which bagging can greatly decrease without increasing the bias by much. For bagging, many samples are drawn from the available data, classification or regression techniques are applied to these perturbations of the original data set to obtain predicted classes for the inputs, followed by combining the results, through averaging for regression and simple voting for classification, to obtain a single classification or regression prediction (Breiman, 1996), referred to as a perturb and combined method by Breiman (1998).

Whereas bagging can successfully be used with simple "off-the-shelf" classifiers (as opposed to classifiers that have been carefully tuned and tweaked) the improved performance does come at the cost of increased computation time and model complexity.

In the classification approach for trees-with-bagging, each base learner casts a vote for the true class of a test sample and the class assignment is made on the aggregate of the votes cast. A database is then collected of the percentage of misclassifications for each base learner. The distribution of classification accuracies can then be inspected for an indication of the relative variable importance using a box plot of this database. The median of this box plot would therefore represent the relative variable importance and the quartiles would give an indication of the confidence in this importance assessment.

With X being the data observations and k the corresponding class labels indicating to which class each observation in X belong, the algorithm for trees-with-bagging can be summarised as follow:

1. From the data observations, X, randomly sample with replacement a training data set, X_{train} and k_{train}, and a test data set, X_{test} and k_{test}.
2. Using X_{train} and k_{train}, train a classification or regression model.
3. Using X_{test} and k_{test}, assess the model performance.
4. Store tree performance and repeat from 1 until a models have been constructed.
5. Repeat from 1, each time omitting a different variable from the data observations

Although such ensembles of tree models are powerful at capturing non-linear relationships between variables, they are black box models that cannot be interpreted directly. These models can, however, be used to generate variable importance measures that give a ranked indication of the relative significance of input variables to the classification (or regression) problem at hand (Auret and Aldrich, 2011).

The hypothesis is that the variable which contributed the greatest to the changes identified will have the greatest influence on the classification accuracy of the models. That is to say that the omission of variables of greater importance should result in models which perform poorer with regards to their

Stellenbosch University http://scholar.sun.ac.za

classification accuracy. For variable importance, the change in median classification accuracy (the median of test observations classified correctly) compared to the models trained with all variables is calculated.

5.1.3 Random Forests

Random forests are an ensemble learning technique (Breiman, 2001) that computes non-linear regression or classification models through the generation and aggregation of multiple regression or classification trees, following an efficient strategy aimed at increasing diversity between the trees (Figure 38). It has been shown that growing these ensembles of classification trees, where each tree in the ensemble depends on a random vector sampled independently from the data, and having them vote for the most popular class have led to significant improvements in classification accuracy.

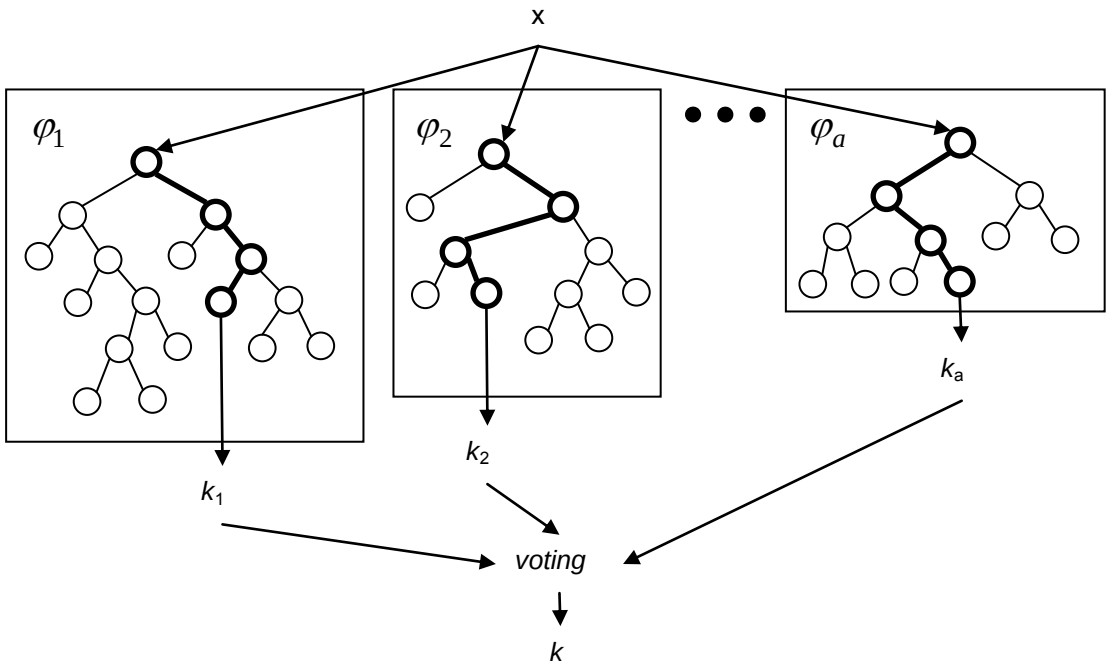

Figure 38: **A general architecture of a random forest where a is the number of trees and k_1, k_2, k_B and k are class labels (adapted from Verikas et al., 2011)**

A classification tree is a hierarchical set of nodes, dividing the feature space into recursive binary partitions, with each node containing a decision involving the comparison of an attribute with a given threshold, allocating a fixed-value prediction to each partition. To evaluate data, each sample is passed down the tree from a root node, through intermediate nodes and ultimately to a terminal node with an associated product result. Classification trees are built from a set of samples using a recursive algorithm. At each step the procedure evaluates all available attributes and possible thresholds, selecting the

combination that maximises a given fitness measure. Following this, the data set is split according to the selected decision and the procedure reapplied to each new subset.

For a classification problem the fraction of samples allocated to the node, η, with class outcome, κ, can be defined as:

$$\hat{p}_{\eta k} = \frac{1}{N_\eta} \sum_{x_i \in R_\eta} I(y_i = \kappa) \tag{5-9}$$

where y is the response vector, x is the predictor vector and R_η and N_η the decision tree local region and number of observations respectively for the node. Furthermore, the class outcome (the majority class in that node) can be defined as $\kappa(\eta) = \arg\max_k \hat{p}_{\eta k}$ with the node impurity measure of tree φ being defined as $Q_\eta = \sum_{\kappa=1}^{C} \hat{p}_{\eta\kappa}(1 - \hat{p}_{\eta\kappa})$, where Q_η is the Gini index often used in classification problems.

In general, classification trees are known for their transparency and sensitivity to small perturbations in the learning data set. A small change in the data set can result in large changes in the developed model (Breiman, 1996). Fitting the model using bagging is one way of mitigating this sensitivity to small perturbations. Bagging is the selection of a random subset of the available data set of the same length and taken with replacement. Each sample is therefore selected at random (to enhance accuracy) from the available data set irrespective of whether it has been picked before.

An ensemble predictor solves a single problem by combining a set of individual discriminant functions. Depending on the ensemble construction strategy used, individual functions

- can belong to a diverse group of models (such as classification trees) or
- can be fitted to a different subset of the full data set (bagging) or
- can differ on the initial conditions of the fitting procedure.

Evaluating the predictor by each discriminant function and then combining the individual outputs to form a final decision, usually using a majority-vote-rule or averaging the predictions, gives a prediction on a new sample. For the majority-vote-rule each function votes for one of the possible outputs with the response of the ensemble given as the one with the most votes. A necessary and sufficient condition for an ensemble of classification trees to be more accurate than any of its individual members is that the members of the ensemble perform better than random and are diverse (Peters et al., 2007).

In addition to resampling of the data with replacement (bagging), for random forests, diversity among the classification trees is further increased through randomly changing the predictive variable sets over the different tree induction processes. Furthermore, whereas standard classification trees are built by splitting each node using the best split among all predictive variables, for random forests each

Stellenbosch University http://scholar.sun.ac.za

classification tree is grown using another bagging subset X_O of the original data set X and the nodes are split using the best split predictive variable among a subset of m randomly selected predictive variables. This increased diversity in the model results in easy-to-build ensembles with very good predictive performance. The random forests algorithm comprising of a classification trees can be defined as follow (Peters et al., 2007):

1. First, a bagging subset X_O using approximately $2/3$ of the elements of the original data set X is created (samples are selected with replacement from all available training samples).
2. Next, an unpruned classification tree, φ_a, is grown to the maximum depth using X_O, randomly selecting m predictive variables out of the p available variables and choosing the best split among these variables at each node, η. Usually $m << p$ and it is suggested as a starting point to use either $m = \sqrt{p}$ or $m = \left[\log_2(p)+1\right]$ (Verikas et al., 2011). Breiman (2001) has shown that selecting $m = \sqrt{p}$ usually gives near optimal results. Another approach would be to define more features by taking random linear combinations of a number of the input variables (Breiman, 2001). For incommensurable input variables in the data set their means are subtracted and they are divided by the standard deviations (normalised) using the mean and standard deviations from the training data.
3. Repeat from step 1 until a classification trees have been constructed.
4. Using either the majority vote or averaging the predictions of the ensemble of a classification trees, predict new data.

An unbiased estimate of the generalisation error can be obtained during the construction of a random forest by considering the elements not used when creating the bagging subset X_O. These out-of-bag (OOB) elements are not used in the construction of the i th tree and therefore on average each element of the original data set X is OOB in $1/3$ of the a tree constructing iterations. The percentage of misclassifications over all OOB elements is called the OOB error which in turn is an unbiased estimate of the generalisation error: the lower the error the stronger the classifier. Using the OOB error estimates eliminates the need to set aside a test data set. It has been shown (Breiman, 2001) that given a sufficiently high number of trees, a, the generalisation error will always converge, resulting in random forests never over-fitting. Also, as the number of combinations increases the error rate decreases, resulting in the OOB estimates overestimating the current error rate. To overcome this and produce unbiased OOB estimates, it is necessary to run past the point where the test data set error converges. Breiman (2001) further showed that an upper bound of the generalisation error can be derived by considering the strength of each individual tree in the forest (a measure of the accuracy of the individual classification trees) and the correlation between any two trees in the forest (a measure of the diversity of the different classification trees): the combination of these measures determining the accuracy of the random forest. Furthermore, the number of predictive variables (features) used to split the nodes, m, is

Stellenbosch University http://scholar.sun.ac.za

directly related to both the strength (higher is better) and correlation (lower is better) of the random forest: reducing m decreased both strength and correlation.

Random forests for feature extraction can be constructed using unsupervised learning techniques. For unsupervised learning the data set used for constructing the random forest model consists of a data set without any class labels or response variables. For this, a proximity matrix needs to be calculated from the data where the random forest dissimilarity measure is the inverse of the proximity matrix (Auret and Aldrich, 2010):

1. First a synthetic data set, X_0, is created from the original, unlabelled data set, X, by randomly sampling from the product of the marginal distributions of X. This results in X_0 having a distribution of independent random variables with corresponding variables between the X and X_0 data sets having the same distribution while destroying the dependency structure in the original data set for the X_0 data set.
2. Next, X is labelled as class 0, X_0 is labelled as class 1 and $Z = \left[X; X_0\right]$ is constructed.
3. For the prediction of the $(0,1)$ labels of the samples in Z, construct a random forest classifier on Z.
4. Initialise a null matrix $S = 0$ with the number of rows and columns equal to the number of samples in X.
5. For each decision tree in the random forest classifier, determine whether the data point pair combinations x_i and x_j, where $i, j = 1,2,\ldots,N$, report to the same terminal node of the tree. If they report to the same terminal node, increase s_{ij} by 1, where s_{ij} is the $(i, j)^{th}$ element of S. Repeat for all trees and all data point pairings.
6. Scale S by dividing by the number of trees in the forest with the resulting similarity matrix S being symmetric, positive definite and with entries ranging between 0 and 1.
7. Construct the dissimilarity matrix $D = I\text{-}S$.
8. Lastly, calculate the multidimensional scaling coordinates in a dimensions. This is done by preserving the pair-wise dissimilarities of the original data points (d_{ij}) as Euclidian distances of the new feature vectors t in the mapping (δ_{ij}), where the mapping can be induced by minimising the cost function $E = \sum_{ij} \left(\delta_{ij} - d_{ij}\right)^2$.

Advantages of random forests include the fact that no assumptions need to be made regarding the nature or distribution of the time series and it can deal with complex non-linear dynamic systems (Auret and Aldrich, 2010). Furthermore, Auret and Aldrich (2010) found that the random forest approach may afford

Stellenbosch University http://scholar.sun.ac.za

distinct advantages over other linear equivalents, particularly when complex non-linear systems need to be monitored.

For systems where the number of candidate predictors, x, is larger than the number of observations, n, traditional statistical modelling approaches are often not estimable while also typically assuming independence among the candidate predictors. With x significantly larger than n, various models, most of which have comparable error estimates, are normally possible, however, each potentially providing a different interpretation of the data. For these cases, machine learning approaches for class prediction are commonly applied (Archer & Kimes, 2008).

Whereas machine learning methods such as linear discriminant analysis require the predictor space to be reduced prior to modelling, random forests do not have this requirement. Compared to the trees-with-bagging approach, random forests for variable importance include a random split algorithm at each node in the tree and does not have any tree pruning or stopping rules. The inclusion of a random split algorithm at each node in the tree allows a large number of substantially different models to be generated, decreasing the correlation between the trees in the forest thus decreasing the forest error rate. Whereas classification random forests are used in the algorithm to determine between-class variable importance, replacing it with regression random forests, within-class variable importance can be calculated.

For variable importance, for every bagging sample during training, all the variables and the leave-one-out algorithm is applied on the OOB data. First, all variables are used to predict the class of each observation using the tree model built for the specific bagging sample, calculating the sum of the number of correct predictions. Next, the leave-one-out algorithm is applied by removing a variable from the data set and using the same tree to predict the class of the observations. This is repeated for all the independent variables, removing a single, different variable for each repetition. The number of correct votes for each permuted sample is then summed and subtracted from the prediction using all the variables with the average difference in accuracy between the original sample and the permuted sample over all the tree models indicating the relative variable importance based on an increase or decrease in classification accuracy (Auret and Aldrich, 2011):

$$\omega_i(\varphi_L) = a\left(\varphi_{L(\theta)}\right) - a\left(\varphi_{L^i_{OOB}(\theta)}\right) \tag{5-10}$$

where $\omega_i(\varphi_L)$ is the variable importance measure based on permutation, a is the accuracy of the model and $L^i_{OOB}(\theta)$ is the OOB learning samples with variable i permuted. The removal of the independent variable resulting in the poorest performing classifier should be the variable contributing the most to the identified changes in the system behaviour. This approach is the most widely used score of importance of a given variable, also being a more reliable indicator when compared to the random forests Gini importance (Genuer et al., 2010). For ensembles of trees the individual tree importance measures can be averaged:

$$\omega_i = \frac{1}{a}\sum_{k=1}^{a}\omega_i\left(\varphi_{L(\theta_k)}\right) \tag{5-11}$$

Archer & Kimes (2008) have found that the random forest methodology is effective for classification problems where the aim is to produce an accurate classifier that can provide insight regarding the discriminative ability of individual predictor variables. The following interesting properties of random forest variable importance measures should also be noted (Auret and Aldrich, 2011):

1. It has been shown that variable importance measures based on permutation are sensitive to the correlation structure of the predictor variables. During permutation of variable X_i, the association between X_i and the response y is destroyed. However, this also destroys the association between X_i and the other input variables $X_j\ (j \neq i)$. A decrease in model accuracy can then not only imply a dependence of y on X_i, but also a dependence of X_i on any of $X_j\ (j \neq i)$, the latter not being useful in the context of variable importance measures.
2. Variable importance measures such as ω_i is not of an univariate nature, but considers multivariate interactions with other input variables.
3. It has been found that the optimal random split parameter m for model accuracy is not necessarily optimal for variable importance purposes. Whereas larger values of m increase the magnitude of variable importance for truly important variables, smaller values of m may allow correlated variables higher importance than merited.

5.1.4 ELM-with-Bagging

Similar to the statistical data-based fault detection, extreme learning machine can be a viable alternative to the CART algorithm used in the trees-with-bagging variable importance analysis (based on the method for developing bagging predictors (Breiman, 1996)). Whereas Brown et al. (1993) showed that neural network models (and therefore ELM) can outperform CART models on multimodal classification problems of large data sets with few attributes, Razi and Athappilly (2005) found that when predicting a continuous variables from categorical variables, neither neural network nor CART models showed a clear advantage of one over the other when considering prediction accuracy. Neural network methods do, however, have the advantage of good generalization, being able to not only handle problems with many parameters well but also classify objects even when the distribution of the objects is very complex. For the sake of completeness, the overview of ELM from section 3.4.4 is restated here.

ELM is a learning algorithm for single hidden layer feedforward neural networks (SLFNs) where the input weights are chosen randomly and the output weights determined analytically (Huang et al., 2006). Whereas traditional learning algorithms employed by feedforward neural networks are limited in their learning speed, mainly due to the slow gradient descent-based nature of the algorithms and the iterative parameter estimation, ELM can not only achieve extremely fast learning speed but also tends to provide

Stellenbosch University http://scholar.sun.ac.za

good generalization performance (thanks to the learning algorithm not only typically reaching the smallest training error but also the smallest norm of weights). Furthermore, for conventional feedforward neural networks there exists a dependency between the different layers of weight and bias parameters which requires all parameters to be tuned. For SLFNs with N hidden nodes it has been shown that with randomly chosen input weights and hidden layer biases exactly N distinct observations can be learned (Huang, 2003), limiting the aforementioned dependency and significantly reducing the number of parameters that needs to be estimated. This allows such an SLFN to be considered as a linear system with the output weights being analytically determined through simple generalized inverse operation of the hidden layer output matrices (Huang et al., 2006), also making it ideal for use when analyzing very large data sets.

Given a training data set consisting of N arbitrary distinct samples (x_j, y_j), a hidden node number $\tilde{N}$, and an activation function $g(x)$, the ELM algorithm can be defined as (Huang et al., 2006):

1. Randomly assign the input weights z_j and biases b_j, $j = 1,\ldots,\tilde{N}$.
2. Calculate the hidden layer output matrix P.
3. Calculate the output weights $\beta = P^H Y$ where $Y = [y_1,\ldots,y_N]^T$

In the classification approach for ELM, a 1-of-k output encoding (one node for each class) is used with the result that the number of output neurons for the SLFN is set equal to the number of classes and the classes relabelled e.g. for a 2 class classification problem the class labels become $[1 \quad -1]$ and $[-1 \quad 1]$. Each base learner casts a vote for the true class of a test sample and the class assignment is made on the aggregate of the votes cast. A database is then collected of the percentage of misclassifications for each base learner. The distribution of classification accuracies can then be inspected for an indication of the relative variable importance using a box plot of this database. The median of this box plot would therefore represent the relative variable importance and the quartiles would give an indication of the confidence in this importance assessment.

With X being the data observations and k the corresponding class labels indicating to which class each observation in X belong, the algorithm for ELM-with-bagging can be summarised as follow:

1. From the data observations, X, randomly sample with replacement a training data set, X_{train} and k_{train}, and a test data set, X_{test} and k_{test}.
2. Using X_{train} and k_{train}, train a classification or regression model.
3. Using X_{test} and k_{test}, assess the model performance.
4. Store tree performance and repeat from 1 until 200 models have been constructed.
5. Repeat from 1, each time omitting a different variable from the data observations

Stellenbosch University http://scholar.sun.ac.za

For variable importance, the hypothesis is that the variable which contributed the greatest to the changes identified will have the greatest influence on the classification accuracy of the models. That is to say that the omission of variables of greater importance should result in models which perform poorer with regards to their classification accuracy. For this, the change in median classification accuracy (the median of test observations classified correctly) compared to the models trained with all variables is calculated.

5.2 Interactive variable importance

5.2.1 Biplots and alpha bags

Biplots, introduced by Gabriel (1971), can be described as a means with which to visualise the scores and loadings obtained from applying SVD to a data matrix. By applying different data pre-processing routines to the data matrix, different types of biplots can be constructed. Gower and Hand (1996) regarded biplots as multivariate analogues of scatterplots, relating the plotted points to the original variables through the biplot axes. This not only increases the ease of interpretation of biplots, but also makes them accessible to non-statistical audiences.

Consider a data matrix X of size $n \times p$. For the Gower and Hand biplot, using the PCA method, the observations are projected onto a r-dimensional subspace O that best fits the observations in a least squares sense (the sum of squared distances from the observations to their respective projections being a minimum) such that:

$$\left\| X - \hat{X} \right\| = trace\left[\left(X - \hat{X} \right)^T \left(X - \hat{X} \right) \right] \qquad (5\text{-}12)$$

for all matrices $\hat{X}$ of rank at most r. The subspace of best fit is found as having the minimum sum of squared distances from the sample points to their projections following the projection of the p-dimensional observations onto the subspace O.

The principal components of the data matrix subsequently form an orthonormal basis for the subspace, allowing these principal axes to be used as scaffolding for plotting the projections of the observations in the biplot. The scaffolding axes can also be replaced by calibrated interpolation or calibrated prediction axes (Gower and Hand, 1996). Whereas interpolation axes are used to find the subspace values of new observations, prediction axes are used to infer the values of the original observations from the subspace values. Both interpolation and prediction axes can be calibrated in the original scales of measurement allowing visual inspection to be used for reading off the values of variables. In addition to this, the direction of the biplot axes gives an indication of the correlation between the variables (although

correlations are not approximated directly), a small angle between two biplot axes suggesting a high correlation between the corresponding variables.

The overall quality of the display can be measured by the ratio:

$$\left(\sum_{i=1}^{a}\lambda_i\right)\Big/\left(\sum_{i=1}^{p}\lambda_i\right) \qquad (5\text{-}13)$$

where λ_i is the eigenvalue associated with the ith principal component, p is the number of variables and a is number of principal components selected for the biplot. For a 2-dimensional biplot this reduces to the amount of total variance represented in the two leading PCA scores. A low quality means that the total structure or information in the data is not well represented by the model. However, any clustering or structure observed in the low dimensional representation also hold for higher dimensions, but there might be clustering or structure in the complete space that cannot be appreciated in a reduced space. Furthermore, the quality of the display of each of the original p variables can also be measured as the adequacy of representation. This adequacy of the representation for the ith variable is given by the ith diagonal element of the $p \times p$ diagonal matrix $diag\left(l_a l_a^T\right)$ (Gower and Hand, 1996), where l is the eigvector matrix. A better measure of accuracy of the individual biplot axes, termed axes predictivities (Gardner-Lubbe, 2008), in reproducing the data can be determined by the degree to which the columns of the approximation, $\hat{X}$, agree with the corresponding columns of the original data matrix, X. Axes predictivities are defined when the sum of the squares of the fitted values for each variable is expressed as a ratio to the total sum of squares. Whereas the adequacy of representation has more to do with the disposition of coordinate axes and any related distortions of the scale of each axis, axes predictivities give an immediate estimate of the success in predicting the values taken by the variables associated with X.

Since the PCA biplot does not maximally separate classes, the biplot methodology has been extended to canonical variate analysis (CVA) biplots for analysing data where it is important to distinguish between various classes of multidimensional observations (Gardner, 2001). The aim of CVA is to find the linear combination of predictor variables that will maximise the ratio of the between class to within class variance, the methodology being closely related to linear discriminant analysis (LDA).

Both PCA and CVA biplots can be equipped with classification regions through the use of alpha bags, a derivative of bagplots, to provide for the visual appraisal of the degree of overlap between classes in the data. Bagplots are bivariate generalisations of the box plot where the univariate rank concept is generalised to the concept of the halfspace location depth of a point relative to a bivariate data set (Rousseeuw et al., 1999). The halfspace location depth can be visualised in two dimensions by extending a line beyond the maximum and minimum values of two variables drawn through a point θ.

Stellenbosch University http://scholar.sun.ac.za

The halfspace location depth is subsequently the smallest count of data points on the same side of the line at each rotated position of the line through 360^{o}.

Using biplots, Aldrich et al. (2004) constructed both platinum and copper flotation process charts based on flotation image features. In both cases CVA biplots were used to successfully distinguish between the various froth classes. Gardner et al. (2005) used PCA biplots to identify different operating regimes in a phosphate flotation plant and CVA biplots to determine process conditions differentiating between the desirable and less desirable operating regimes. CVA biplots were also successfully used for the visualisation of the hydrolysis of zinc chloride in ammonium chloride solutions allowing for the classification of different zinc precipitates based on process conditions. Mixed precipitates were then assigned to the different classes by mapping them onto the biplots.

5.3 Technique evaluation

For the evaluation of the variable importance techniques, the 3 data sets / models as described in section 3.5 were used:

- Simple multivariate time series data
- Simple multivariate process (Ku et al., 1995)
- Tennessee Eastman process (Downs and Vogel, 1993)

The evaluation not only focusses on validation of the techniques, but also allows for an appreciation of the techniques and their applicability to be obtained and a comparison to be drawn with the results achieved by the statistical data-based fault detection techniques.

The variable importance performance metrics were subsequently tested on both the NOC data as well as all the predetermined fault condition data for each of these data sets / models. Listed in Table 10 is an overview of all the performance metrics tested.

As required by the CVA biplots technique, data were centred to zero mean prior to analysis. Furthermore, prior to evaluation the variables for each data set were checked to see whether or not the data are independent and identically distributed and have a normal distribution. To test for iid data, each variable was checked to determine if there was any significant autocorrelation in the time series. Significant levels of autocorrelation are indicative of data that is not independent and identically distributed. To test if the data is normally distributed, each variable was subjected to the Lilliefors test. The Lilliefors test is a goodness-of-fit test of composite normality, testing whether or not the data in a time series come from an unspecified normal distribution. Such tests determining the statistical characteristics of the data is critical in that it allows appropriate techniques, having specific assumptions, to be matched up with the data.

In order to effectively compare the variable importance performance metrics they were each evaluated based on their classification error. The classification error, together with its upper confidence limit, for the NOC data, called henceforth the detection threshold (effectively the false alarm rate of the technique), was used as the baseline where no variables were important. For the results interpretation all

unimportant variables are displayed in blue (values below detection threshold), important variables in green (values above the detection threshold) and important variables that are significantly different from any other variables are displayed in red accompanied by the actual classification error as a value text.

For evaluating the CVA biplot variable importance results, the graph displays the axes predictivities for each variable as the marker colour accompanied by the adequacy as a value text. Axes predictivities give an immediate estimate of the success in predicting the values taken by the variables associated with the data. For the PCA SPE variable contributions, the maximum variable contribution value is indicated next to the marker, with the marker colour being a scaled value of this.

Table 10: **Overview of variable importance analysis techniques**

Technique	Type	Metric	Comments
Linear discriminant analysis	Linear	Classification error	Assume normally distributed classes or equal class covariance. Optimal at maximising separation between classes. For regression, collinearity is a problem.
Trees-with-bagging	Non-linear	Classification error	No assumption regarding data distribution. Robust against outliers. Robust against over-fitting.
Random Forests	Non-linear	Classification error	No assumption regarding data distribution. Robust against correlated independent variables. Unreliable when variables are multi-collinear.
ELM-with-bagging	Non-linear	Classification error	No assumption regarding data distribution. Extremely fast learning speed. Collinearity potentially a problem.
CVA biplots	Linear	Visually, axis predictivity and adequacy	Powerful data visualisation. Assume equal class covariance. Graph can become overly cluttered when analysing too many variables simultaneously.
PCA SPE	Linear	SPE	Assume data to be normally distributed.

5.3.1 Simple multivariate time series data

For the simple multivariate time series data 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. For analysis, the confidence limit threshold was set to 99%.

Stellenbosch University http://scholar.sun.ac.za

As previously stated, inspection of the simple multivariate time series data (Figure 10) indicated it to be independent and identically distributed while having a normal distribution. These data characteristics meet the required assumptions of all of the variable importance analysis techniques; ensuring their relevance to the data set being analysed. Furthermore, since the data set comprises entirely of randomly generated data with only a predefined covariance between x_2, x_3 and x_4 it is expected that, although applicable, there would be no need for the additional complexity offered by non-linear techniques. For the sake of completeness, however, all the available algorithms will be assessed. This will not only allow all appropriate algorithms to be evaluated but also help to gain insight into the potential benefits that may be offered by the more complex, non-linear variable importance analysis techniques.

For the step change fault conditions, MTS1-MTS4, all the techniques were able to easily identify the variable that has changed in the data set. For the linear discriminant analysis (Figure 39) and random forest (Figure 41) variable importance measures a few less important variables were also erroneously identified, however, the classification error of these variables were significantly lower than those of the actual important variables, allowing them to easily be discarded as false alarms. Visual inspection of the relevant CVA biplots confirmed their effectiveness.

For the ramp change fault conditions, MTS5-MTS8, all the techniques were again able to easily identify the variable that has changed in the data set. This time, however, the linear discriminant analysis (Figure 39) and trees-with-bagging (Figure 40) variable importance measures erroneously identified a few less important variables, however, as before, the classification error of these variables were significantly lower than those of the actual important variables, allowing them to easily be discarded as false alarms. Visual inspection of the relevant CVA biplots confirmed their effectiveness.

For the spike fault conditions, MTS9-MTS12, none of the techniques were able to identify the variable that has changed in the data set. The trees-with-bagging (Figure 40) variable importance measures erroneously identified a few less important variables, however, as before, the classification error of these variables were so low that they were easily discarded as false alarms. The primary reason for this poor performance is the fact that the classification error is measured over a time span of 100 seconds whereas the spike fault condition only last between 1 and 10 seconds. However, visual inspection of the relevant CVA biplots showed that they were able to detect a change in the data set. Whereas for the change in the uncorrelated variable the CVA biplots were unable to correctly identify the variable that has changed, for the change in the correlated variable the CVA biplots were very successful in identifying the correct variable that has changed. However, the correlated x_3 was visually implicated as an important variable for all 4 fault conditions, and although correlations are not approximated directly for CVA biplots, it is suspected that this behaviour is directly related to the correlated nature of the data set.

For the standard deviation change fault conditions, MTS13-MTS16, as with the spike fault conditions, none of the techniques were able to reliably identify the variable that has changed in the data set. Whereas the extreme learning machine (Figure 42) variable importance measure was unable to identify

any of the important variables, both the linear discriminant analysis (Figure 39) and trees-with-bagging (Figure 40) variable importance measures detected only 1 of the 4 fault conditions, with most of the variable importance techniques, including the random forest (Figure 41) variable importance measure, erroneously identifying a less important variable. As with the spike fault conditions, visual inspection of the relevant CVA biplots did show that they were able to detect a change in the data set. As before, CVA biplots were not very reliable in identifying the responsible variable with the correct variable being identified for only 1 of the 4 fault conditions, as with the other variable importance techniques, and for all other fault conditions identifying the correct variable together with an incorrect one.

For the covariance shift fault conditions, MTS17-MTS19, only the major covariance shift was correctly identified by most of the variable importance techniques. The linear discriminant analysis (Figure 39) variable importance measure incorrectly identified the covariance shift in the uncorrelated variable (fault condition MTS17) and only identified the other correlated variables as important for the covariance shift in the correlated variable (fault condition MTS18). Although not clear from the CVA biplot axes predictivities (Figure 43), following visual inspection of the relevant CVA biplots, it was evident that the CVA biplots were very successful in identifying a covariance change in the data set with a fair indication of the variables responsible for the fault condition.

For the function change fault conditions, MTS20-MTS23, the trees-with-bagging (Figure 40) and random forest (Figure 41) variable importance measures were successful at detected 3 of the 4 fault conditions, with only the smallest change in the uncorrelated variable not being identified at all. In contrast to this, the extreme learning machine (Figure 42) variable importance measure was successful at detecting only 1 of the 4 fault conditions with the linear discriminant analysis (Figure 39) variable importance measure not being able to correctly identify any of the important variables, only flagging various false alarms. Visual inspection of the relevant CVA biplots also indicated that this technique was effective in correctly identifying the important variables for the 2 major changes (fault conditions MTS21 and MTS23).

From this evaluation it was found that all the techniques were effective in identifying important variables due to “simple” changes in the data. However, as expected, for the more “complex” changes the covariance structure of the data seemed to be a challenge for all techniques. When compared to the PCA SPE variable contribution results (Figure 44), the variable importance analyses results were generally similar, although not being able to identify the important variables related to the standard deviation change fault conditions. The ELM-with-bagging variable importance measure proved the least reliable (albeit without ever erroneously identify important variables) when considering the more “complex” changes, with the trees-with-bagging and random forests techniques performing the best and very similarly. It can therefore be concluded that for the simple multivariate time series data, being independent and identically distributed with a normal distribution, the trees-with-bagging and random forests variable importance analysis techniques performed the best of the techniques tested. Additionally, CVA biplots proved to be invaluable, albeit requiring visual inspection for each assessment, in identifying fault conditions and gaining an understanding of the important variables.

Stellenbosch University http://scholar.sun.ac.za

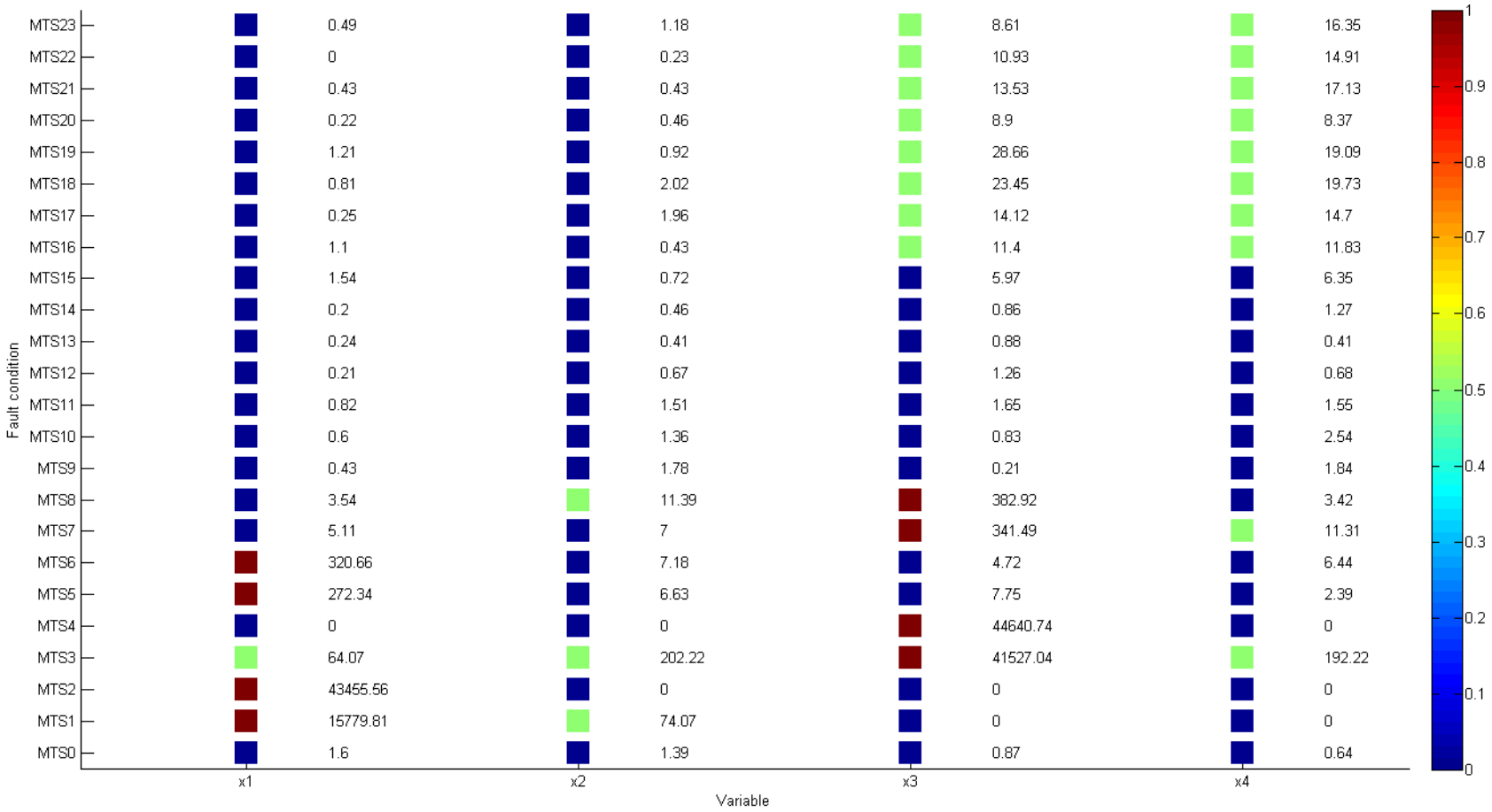

Figure 39: **Simple multivariate time series data: linear discriminant analysis variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

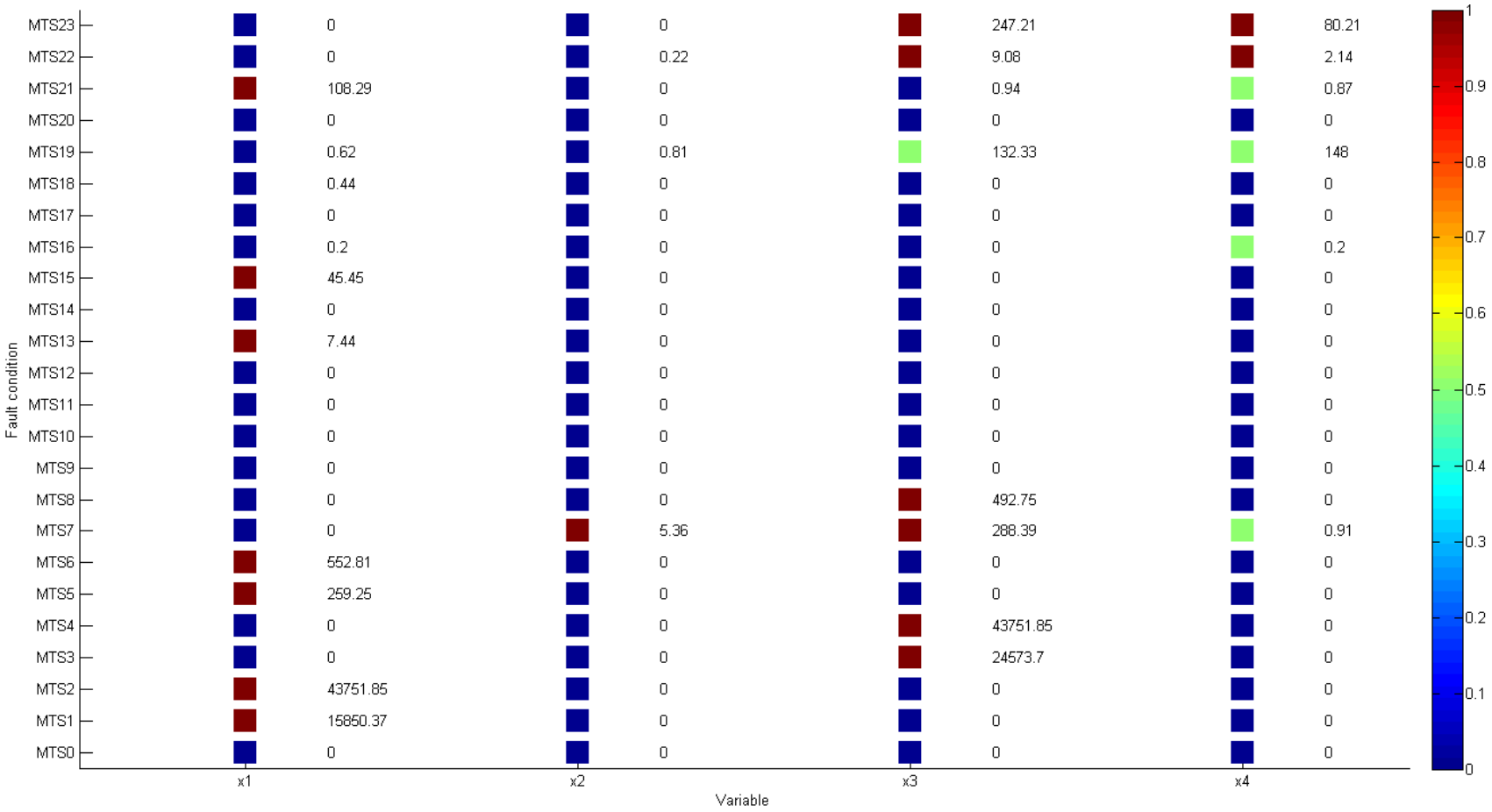

Figure 40: **Simple multivariate time series data: trees-with-bagging variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

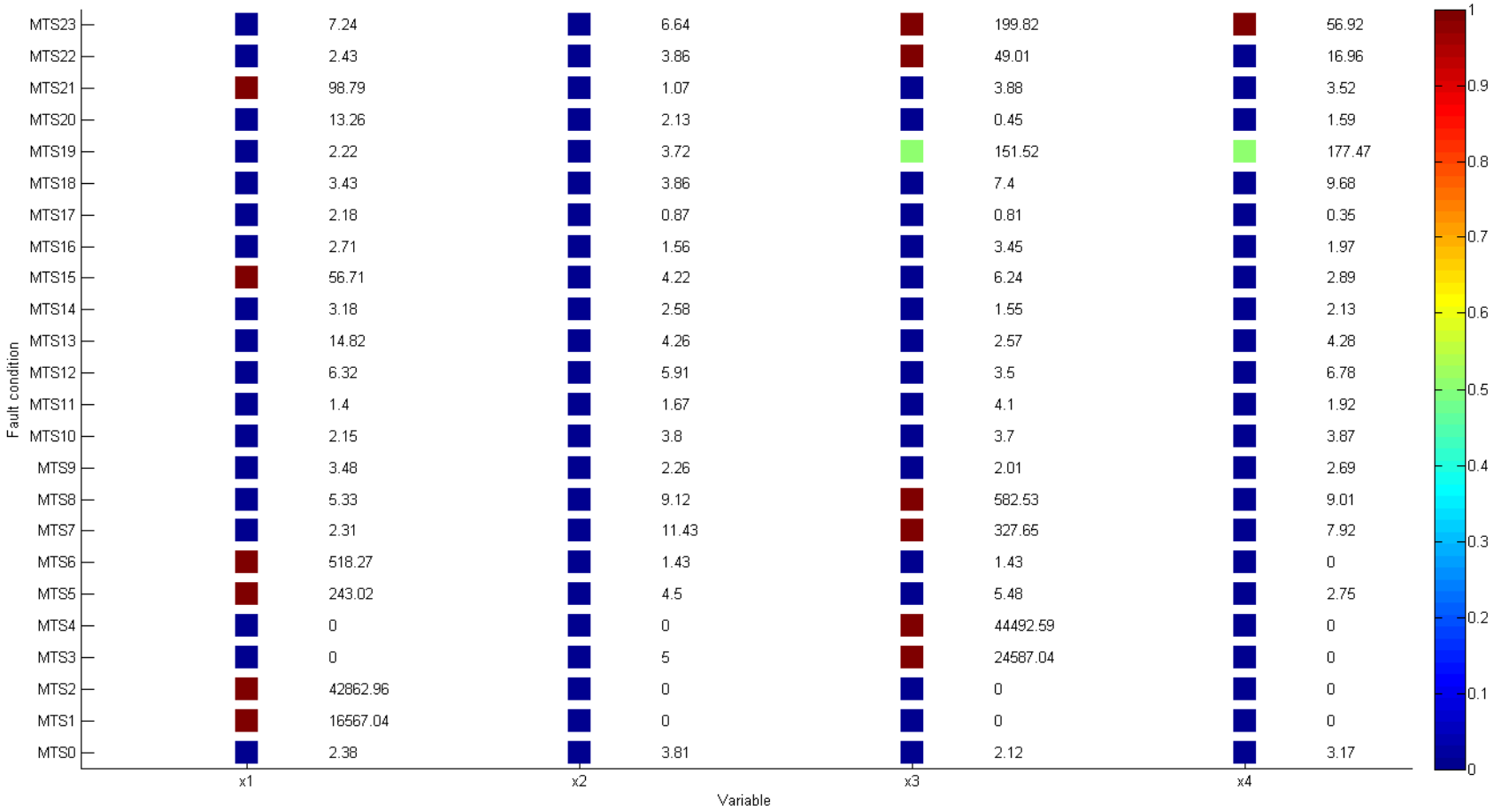

Figure 41: **Simple multivariate time series data: random forests variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

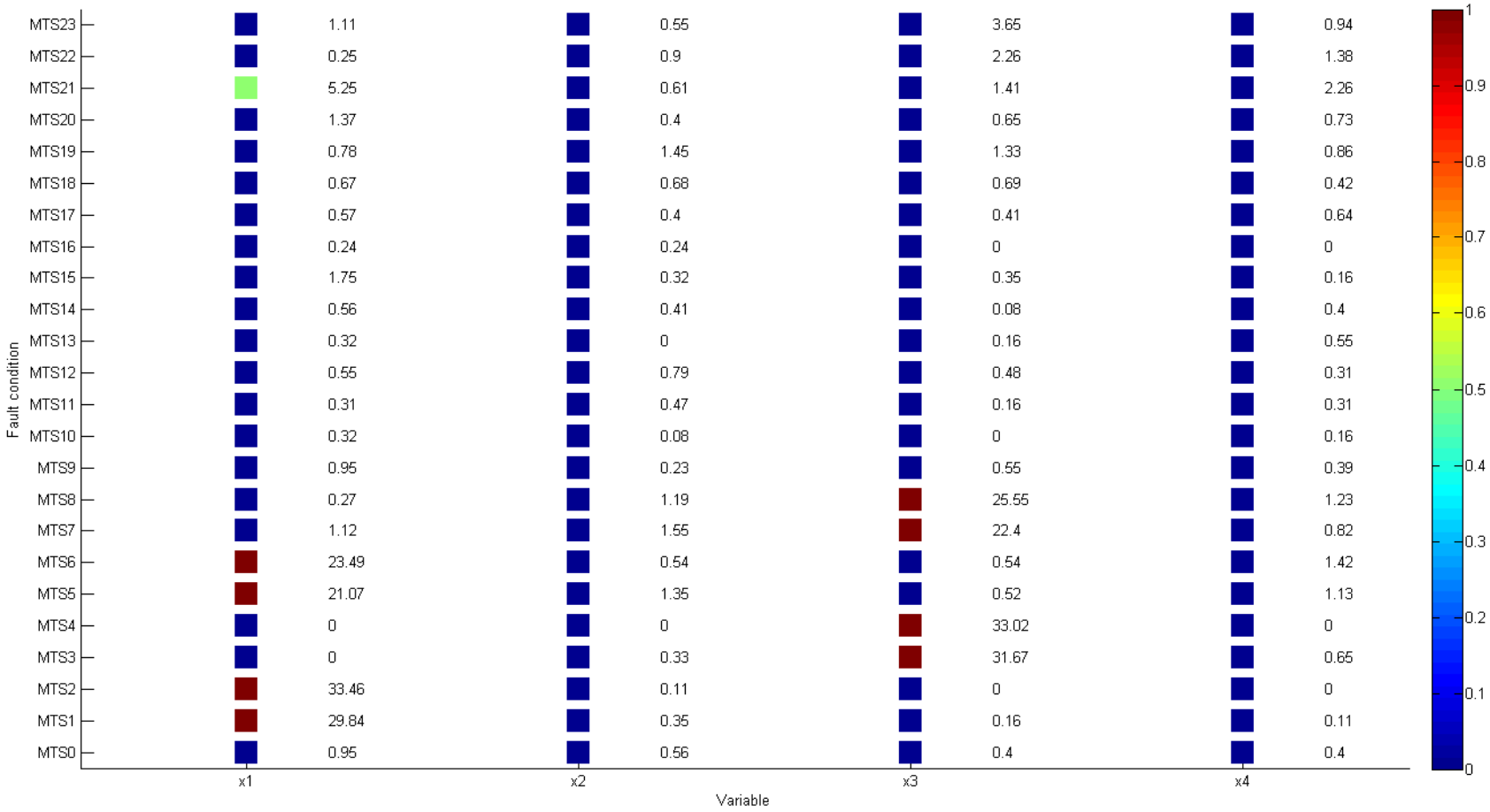

Figure 42: **Simple multivariate time series data: ELM-with-bagging at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

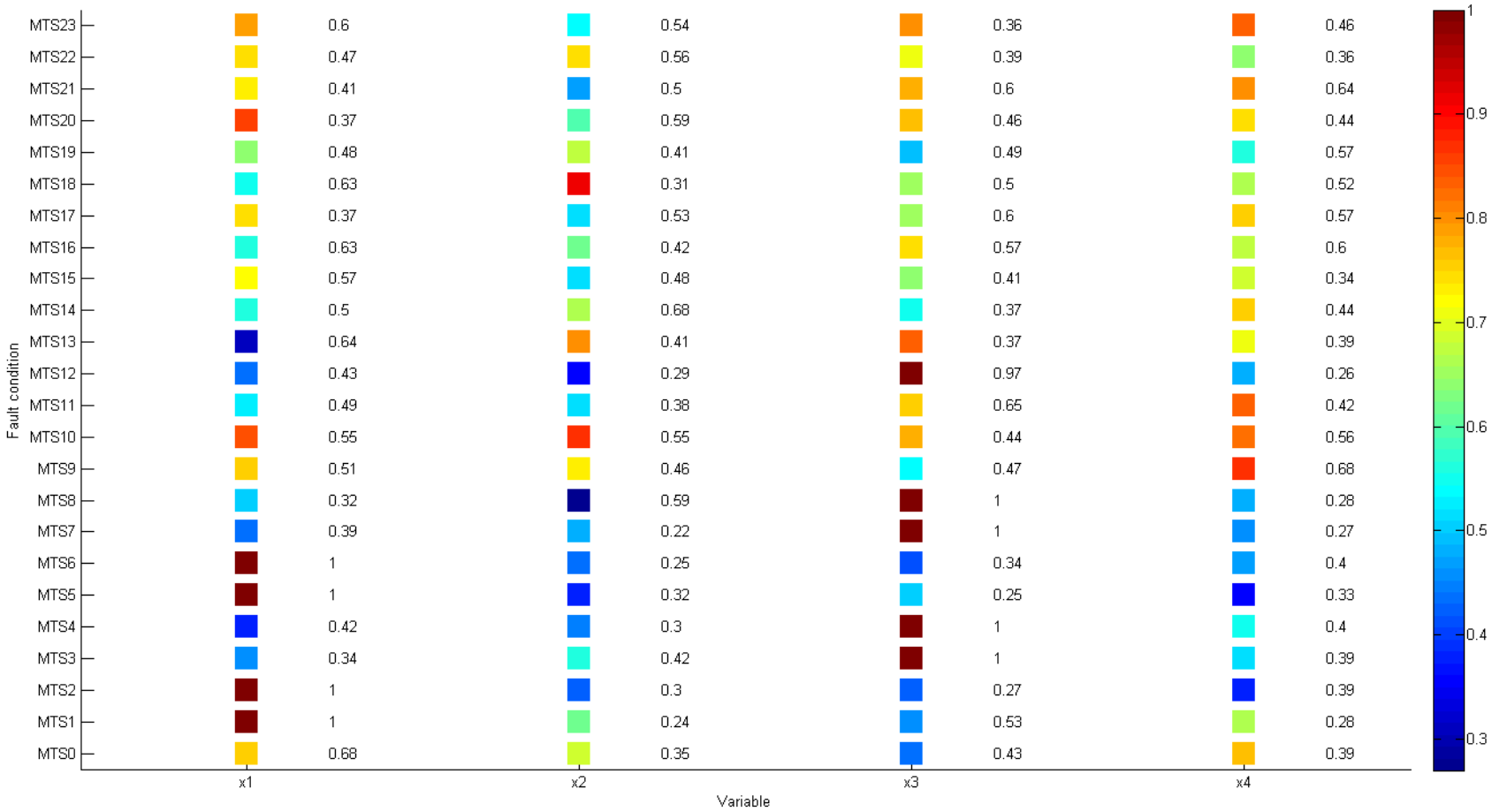

Figure 43: **Simple multivariate time series data: CVA biplot variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

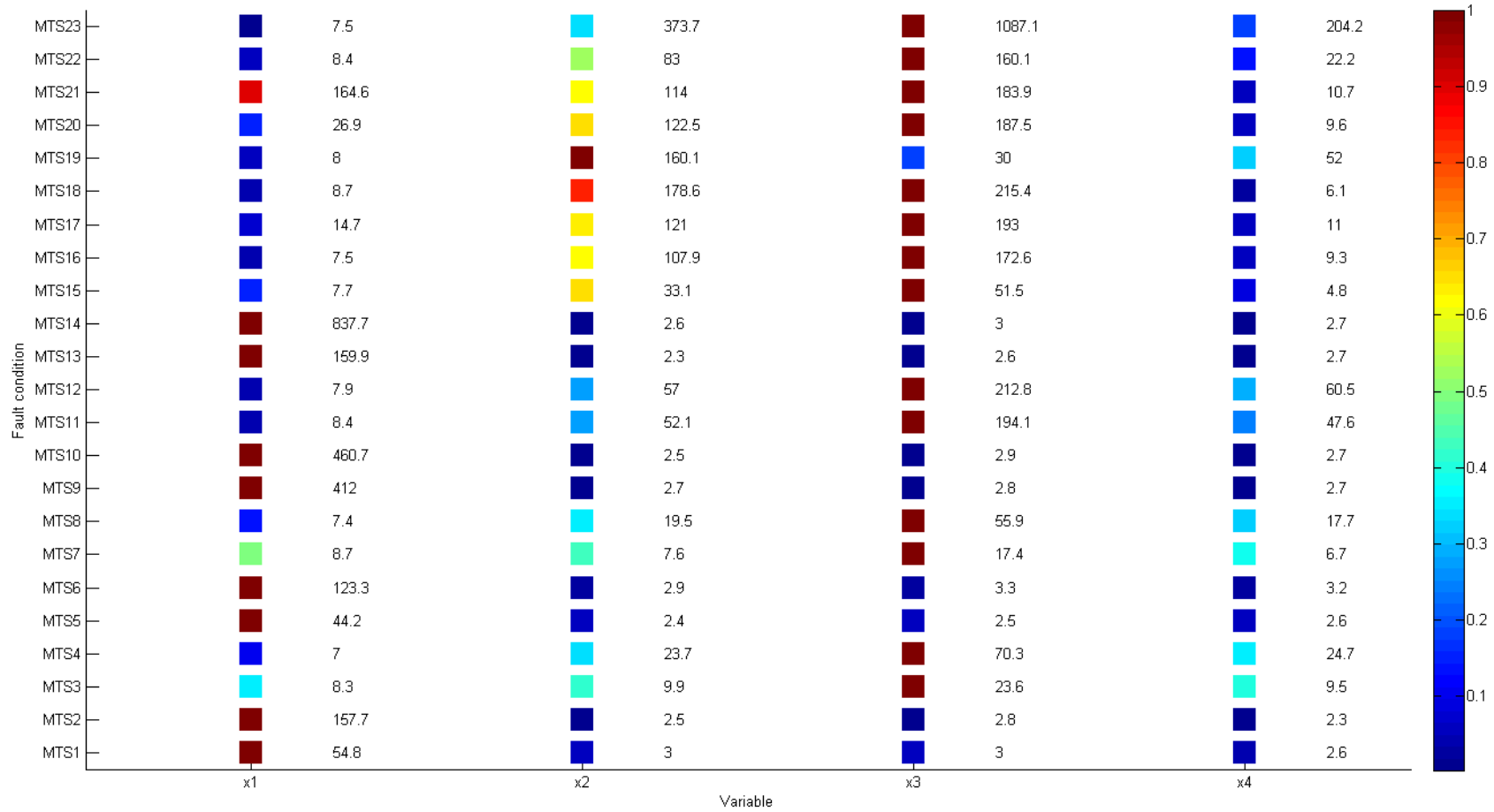

Figure 44: **Simple multivariate time series data: PCA SPE variable contributions at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

5.3.2 Simple multivariate process

For the simple multivariate process 1000 data points at a sampling rate of one sample per second were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. For analysis, the confidence limit threshold was set to 99%.

As stated previously, inspection of the simple multivariate process data (Figure 11) indicated it to not be independent and identically distributed while having a normal distribution. Not only do these data characteristics violate the required assumptions of all of the variable importance analysis techniques, but variable importance analysis techniques typically struggle to correctly predict the most important independent variable in the presence of correlation. However, as stated earlier, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Also, the magnitudes of the process changes resulting in the fault conditions are very small and it is expected that all the available techniques will struggle to reliably identify the important variables. Therefore, for the sake of completeness all the available algorithms will be assessed. In this instance this will help to gain insight into the robustness of the variable importance analysis techniques whose data assumptions have not been met.

For the mean shift fault conditions, SMP1-SMP5, very different results were obtained for the various techniques. The linear discriminant analysis technique (Figure 45) was effective only in detecting the important variables in the outputs of the process, not the inputs. Only for the largest mean shift was the important input variable identified, however, at the cost of the other input variable also being flagged (a false alarm). The trees-with-bagging technique (Figure 46) was very effective in detecting the correct input variable as the most important variable, except for the smallest mean shift, as well as identifying the change in the output variables, clearly distinguishing between the more important input variable and the less important output variables. The random forests technique (Figure 47) performed similar to the trees-with-bagging technique, but only for the two largest mean shifts. The extreme learning machine technique (Figure 48) was effective only in detecting the important variables in the outputs of the process, not the inputs. Only for the largest mean shifts were the important input variable identified, however, at the cost of the other input variable also being flagged (a false alarm). Visual inspection of the relevant CVA biplots confirmed them to be as effective as the random forests technique, but not being able to distinguish between the more important input variable and the less important output variable. In general it was also found that as the magnitude of the mean shift increased, the classification error of the performance metrics increased.

For the parameter change fault conditions, SMP6-SMP8, only the extreme learning machine and trees-with-bagging techniques were able to detect some of the larger of the fault conditions and correctly identify the important variables. Neither the linear discriminant analysis or the random forests nor the CVA biplots technique were able to identify any important variables for any of these fault conditions.

Stellenbosch University http://scholar.sun.ac.za

As expected, due to the small magnitudes of the process changes resulting in the fault conditions, most the available techniques struggled to reliably identify the important variables for all fault conditions. When compared to the PCA SPE variable contribution results (Figure 50), the variable importance analyses performed consistently better due to the fact that the PCA model was unable to reliably detect any of the fault conditions. From the evaluation it can be concluded that for the simple multivariate process data, not being independent and identically distributed with a normal distribution, the trees-with-bagging variable importance analysis technique performed the best of the techniques tested. Although the data characteristics of the data set violated the assumptions of the trees-with-bagging variable importance analysis technique, the technique proved to be exceptionally robust in this regard and was still able to correctly identify the majority of the important variables. With regards to performance, the trees-with-bagging approach was followed by the extreme learning machine approach, with the random forests and CVA biplot techniques only being able to correctly identify the largest fault conditions and the linear discriminant analysis only being effective in identifying the less important variables for the mean shift fault conditions.

Stellenbosch University http://scholar.sun.ac.za

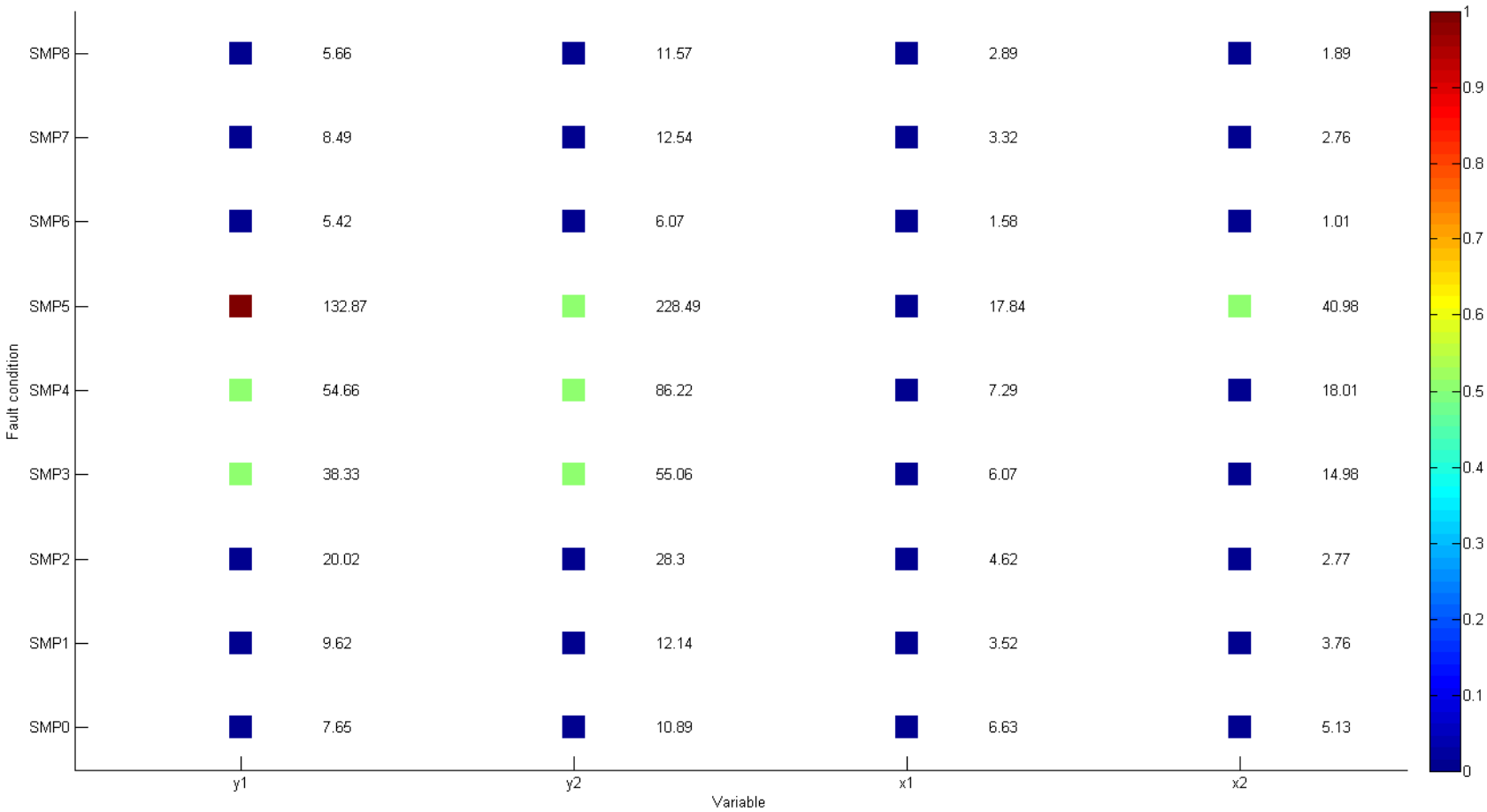

Figure 45: **Simple multivariate process: linear discriminant analysis variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

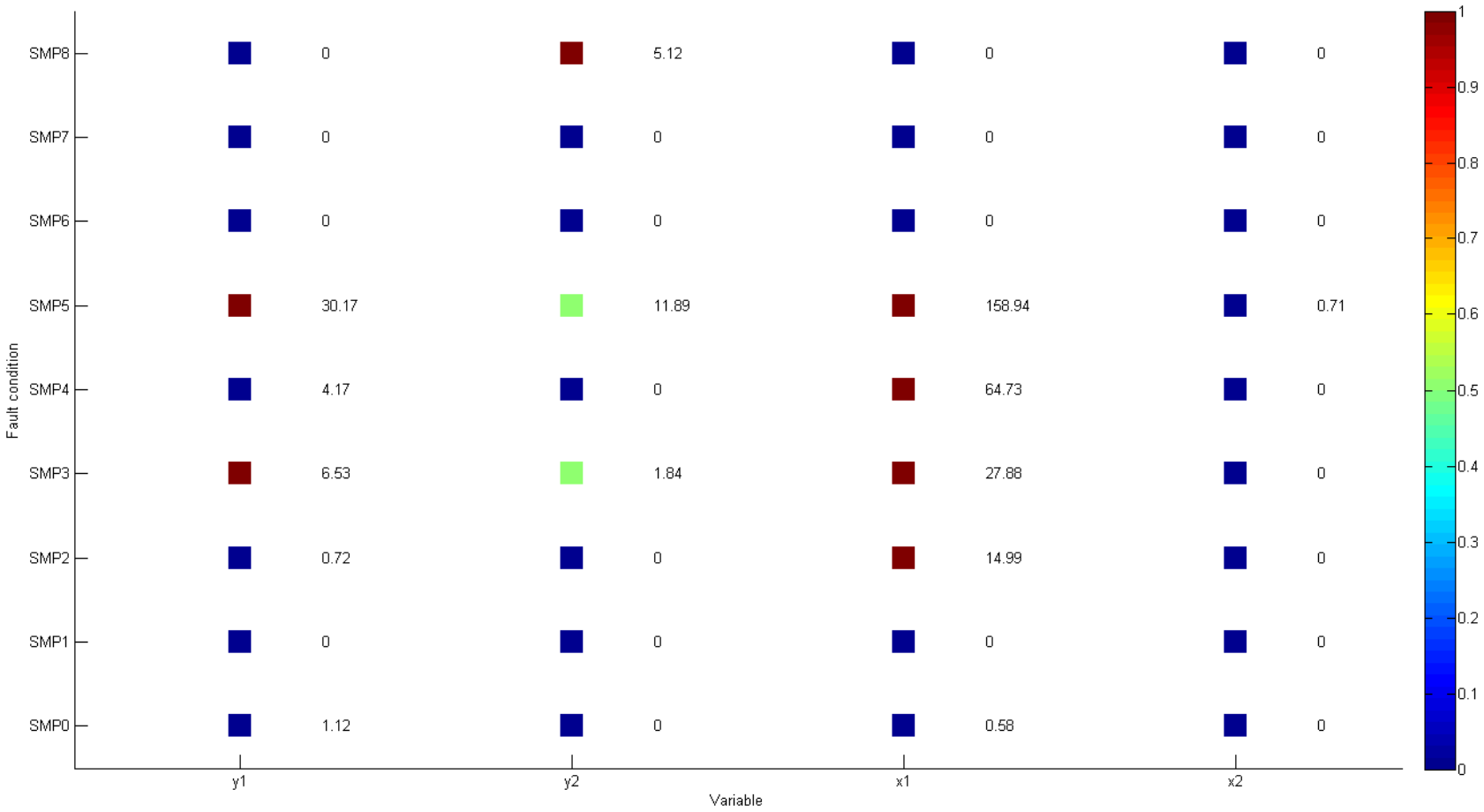

Figure 46: **Simple multivariate process: trees-with-bagging variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

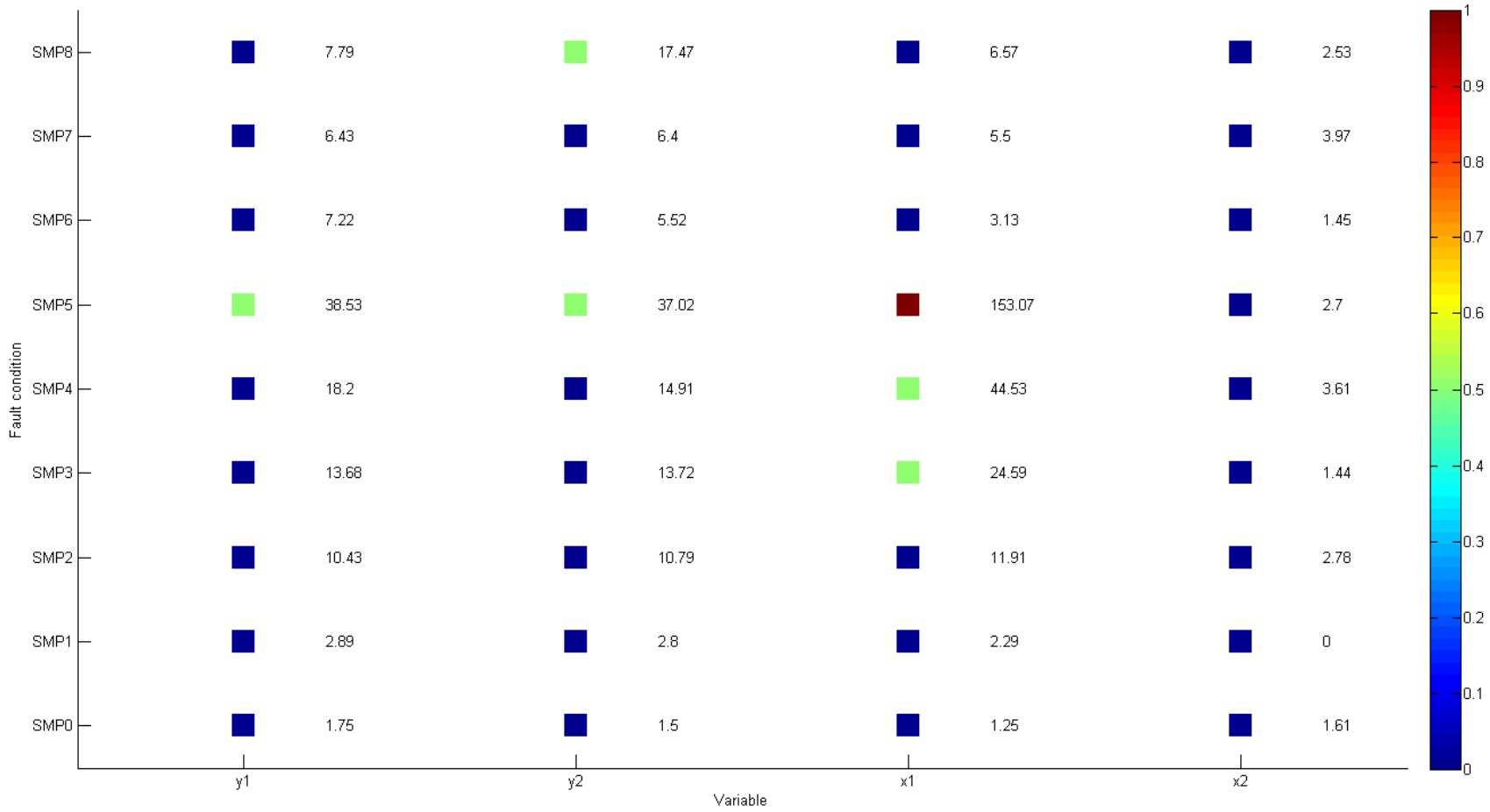

Figure 47: **Simple multivariate process: random forests variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

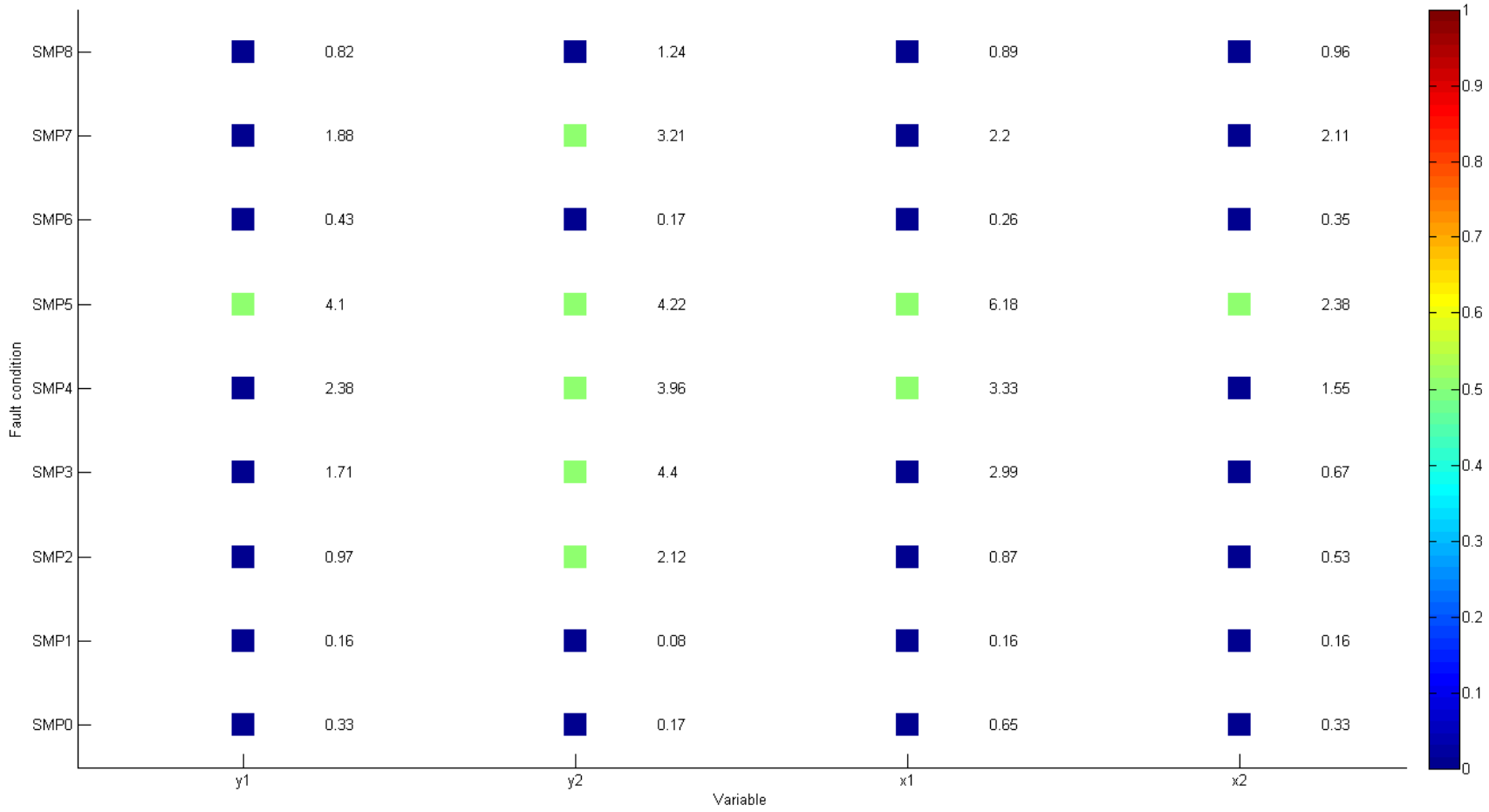

Figure 48: **Simple multivariate process: ELM-with-bagging at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

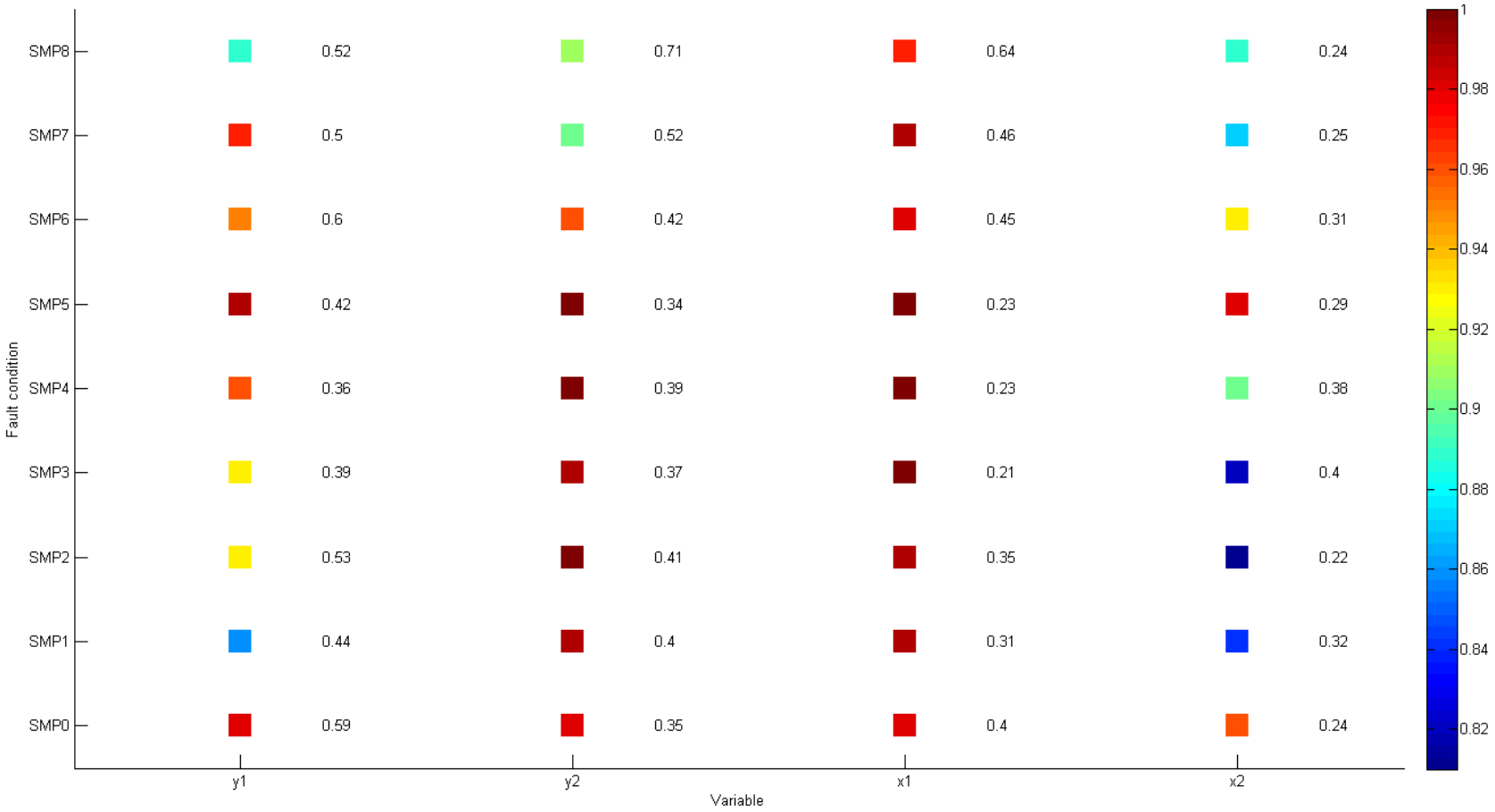

Figure 49: **Simple multivariate process: CVA biplot variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

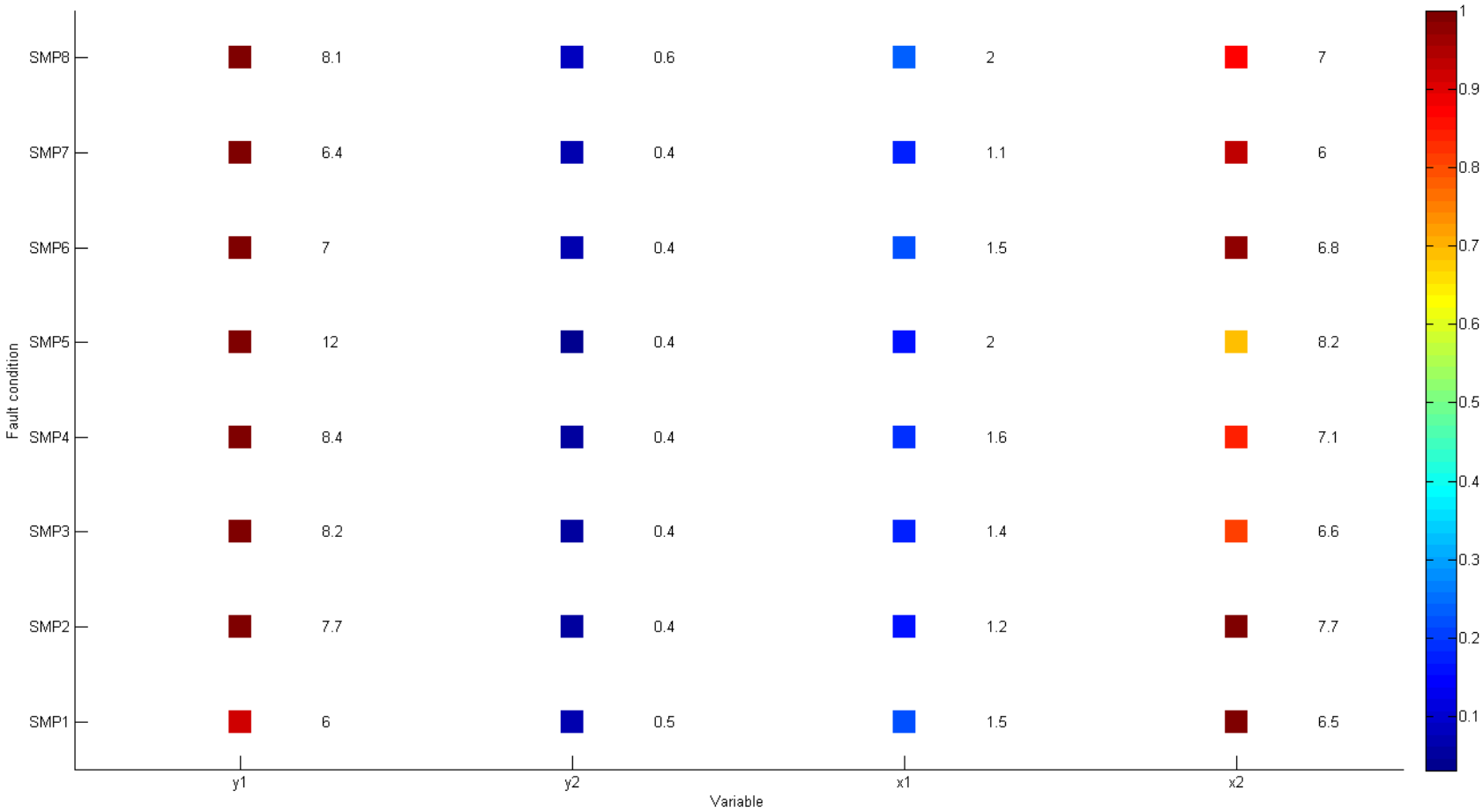

Figure 50: **Simple multivariate process: PCA SPE variable contributions at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

5.3.3 Tennessee Eastman process

For the Tennessee Eastman process 3600 data points at a sampling rate of one sample per ninety seconds were generated for each of the variables with the fault condition introduced at data point 101 and the evaluation criteria determined over the following 100 data points. For determining the reference models the average of 10 data sets were used whereas the evaluation criteria for each fault condition were also based on the average of 10 data sets. For analysis, the confidence limit threshold was set to 99%.

As stated previously, inspection of the Tennessee Eastman process data (Figure 13) indicated it to have a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, with many of the variables having significant autocorrelation with similar profiles. This similarity in the autocorrelation profiles is indicative of the presence of high levels of collinearity (linear correlation structure between the variables). Not only do these data characteristics violate the required assumptions of all of the variable importance analysis techniques, making none of the variable importance analysis techniques suitable for assessing all of the data in this particular data set, but variable importance analysis techniques typically struggle to correctly predict the most important independent variable in the presence of correlation. However, as stated earlier, there is an implicit assumption in the machine learning community that algorithms for which the iid assumptions are violated, will still work well in practice (Dundar et al., 2007). Furthermore, considering that this is quite an extensive data set it is however expected that some of the fault conditions will relate to data that is independent and identically distributed while having a normal distribution, resulting in the other algorithms also being valid some of the time. Therefore, all the available algorithms will be assessed. In this instance this will help to gain insight into the robustness of the variable importance analysis techniques whose data assumptions have not been met.

Since it was found for the Tennessee Eastman process fault detection technique evaluation that not all the variables responsible for the fault conditions or the variables closely related to the fault conditions were being monitored (Chen and McAvoy, 1998), the control system controlling the Tennessee Eastman process was too quick in correcting the induced fault condition, the effect of the fault condition on the process was too slow or the magnitude of the fault condition was too small to result in a measurable process performance degradation over the evaluation period it was decided to determine a techniques success based on whether it could correctly identify a group of important variables, including associated variables as less important.

From the variable importance analysis results (Figure 51, Figure 52, Figure 53, Figure 54 and Figure 55) the simulated fault conditions can be divided into 3 distinct groups:

d) Fault conditions for which the important variables could easily be identified by the majority of the variable importance techniques: fault conditions TEP6, TEP7, TEP10, TEP11, TEP12, TEP21, TEP22, TEP23 and TEP24 (showing some overlap with the fault detection results).

Stellenbosch University http://scholar.sun.ac.za

e) Fault conditions for which the important variables were typically identify by approximately half of the variable importance techniques: fault conditions TEP1, TEP2, TEP3, TEP4, TEP5, TEP8, TEP13, TEP16, TEP17, TEP18 and TEP20 (also showing some overlap with the fault detection results).
f) Fault conditions for which none of the important variables were identify by the variable importance techniques: fault conditions TEP9, TEP14, TEP15 and TEP19 (also showing some overlap with the fault detection results).

The fault conditions in group (a) consist mainly of step changes and a few random variation changes. Important variables for fault conditions TEP6, TEP7, TEP11, TEP12 and TEP24 were correctly identified by all techniques tested. For these fault conditions either the variable responsible for the fault condition or a variable closely related to the fault condition is being monitored. This finding also held for the other fault conditions in group (a).

The fault conditions in group (b) again consist mainly of step changes and "other" disturbances, with a single random variation and ramp change fault condition. Unlike for group (a) fault conditions, for group (b) the variable responsible for the fault condition was rarely directly being monitored, however, variables related to the fault conditions were mostly being monitored. For this group, none of the variable importance techniques stood out as being able to more reliably identify important variables compared to the others.

The fault conditions in group (c) consist of a random variation change and "other" disturbances. For this group, none of the variable importance analysis techniques were able to correctly identify the important variable for any of the fault conditions. As with the Tennessee Eastman process fault detection technique evaluation, this can be ascribed to the fact that variables responsible for the fault conditions or variables closely related to the fault conditions were not being monitored (Chen and McAvoy, 1998), the control system controlling the Tennessee Eastman process was too quick in correcting the induced fault condition, the effect of the fault condition on the process was too slow or the magnitude of the fault condition was too small to result in a measurable process performance degradation over the evaluation period.

From this evaluation it can be concluded that for the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, that very similar performance was achieved by all the variable importance techniques evaluated (with the trees-with-bagging and extreme learning machine techniques performing slightly worse than the others). When compared to the PCA SPE variable contribution results (Figure 56), the variable importance analyses results were generally very similar. It should be noted that for most of the instances where the statistical data-based fault detection techniques were unable to detect the fault conditions, the variable importance analyses were able to identify the important variables.

Stellenbosch University http://scholar.sun.ac.za

Although the data characteristics of the data set violated the assumptions of all of the variable importance analysis techniques, the techniques proved to be fairly robust in this regard and were still able to correctly identify a large portion of the important variables. This less than ideal performance by the techniques being evaluated can in part be ascribed to the fact that variable importance analysis techniques typically struggle to correctly predict the most important independent variable in the presence of correlation. For approximately 40% of the fault conditions all techniques were successful, for another 40% of the fault conditions the techniques were only successful some of the time and for the last 20% of the fault conditions, none of the techniques were successful. A better result could probably have been obtained by monitoring all the variables associated with the Tennessee Eastman process, but in reality this is not always feasible. Due to the complexity of the Tennessee Eastman process and the fact that the results for the technique evaluation of the variable importance measures on this process was done subjectively, it is important to note that an absolute “correct” answer isn’t always obtainable regarding variable importance. In this case the NOC data was used to calibrate the techniques for the various variables. Following the calculation of the performance statistics for the various fault conditions, a “feel” was obtained for the magnitude of the performance statistics. This in turn was used to determine heuristic rules which were used to determine whether or not a specific variable importance technique was successful in identifying the important variables for a specific fault condition.

As for the fault condition technique evaluation, it was again found that the Tennessee Eastman process poses a very complex problem. This was especially evident when viewing the individual CVA biplots generated for each fault condition. Even though the CVA biplot technique can successfully reduce the 16-dimensional problem to 2-dimensions, visualisation on a 2-dimensional plane is a challenge as the graphs become too crowded. It is again suggested that an approach to potentially reducing the complexity inherent to monitoring the Tennessee Eastman process as a whole lies with the idea of simplifying the problem through the use of process causality maps.

Stellenbosch University http://scholar.sun.ac.za

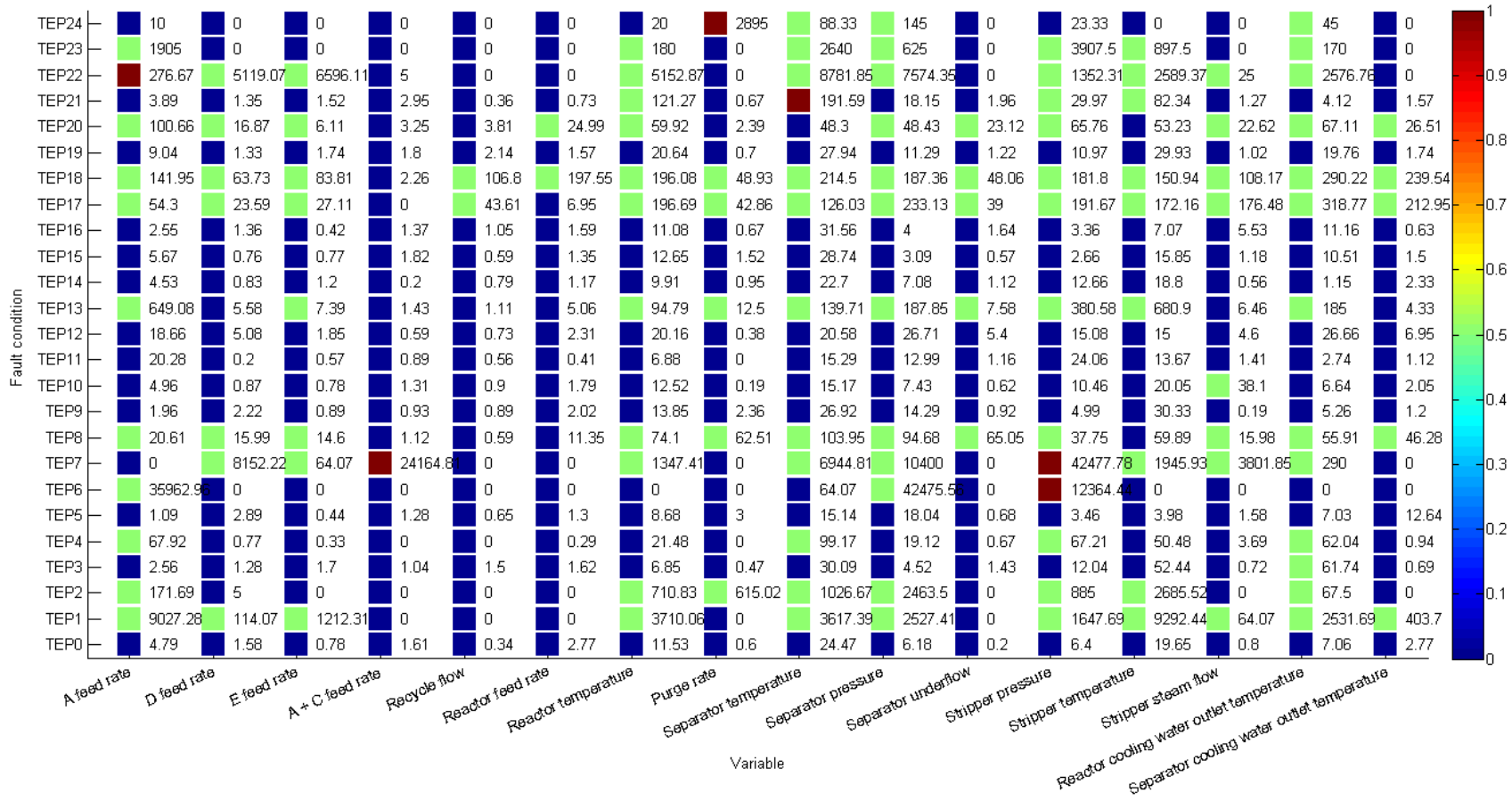

Figure 51: **Tennessee Eastman process: linear discriminant analysis variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

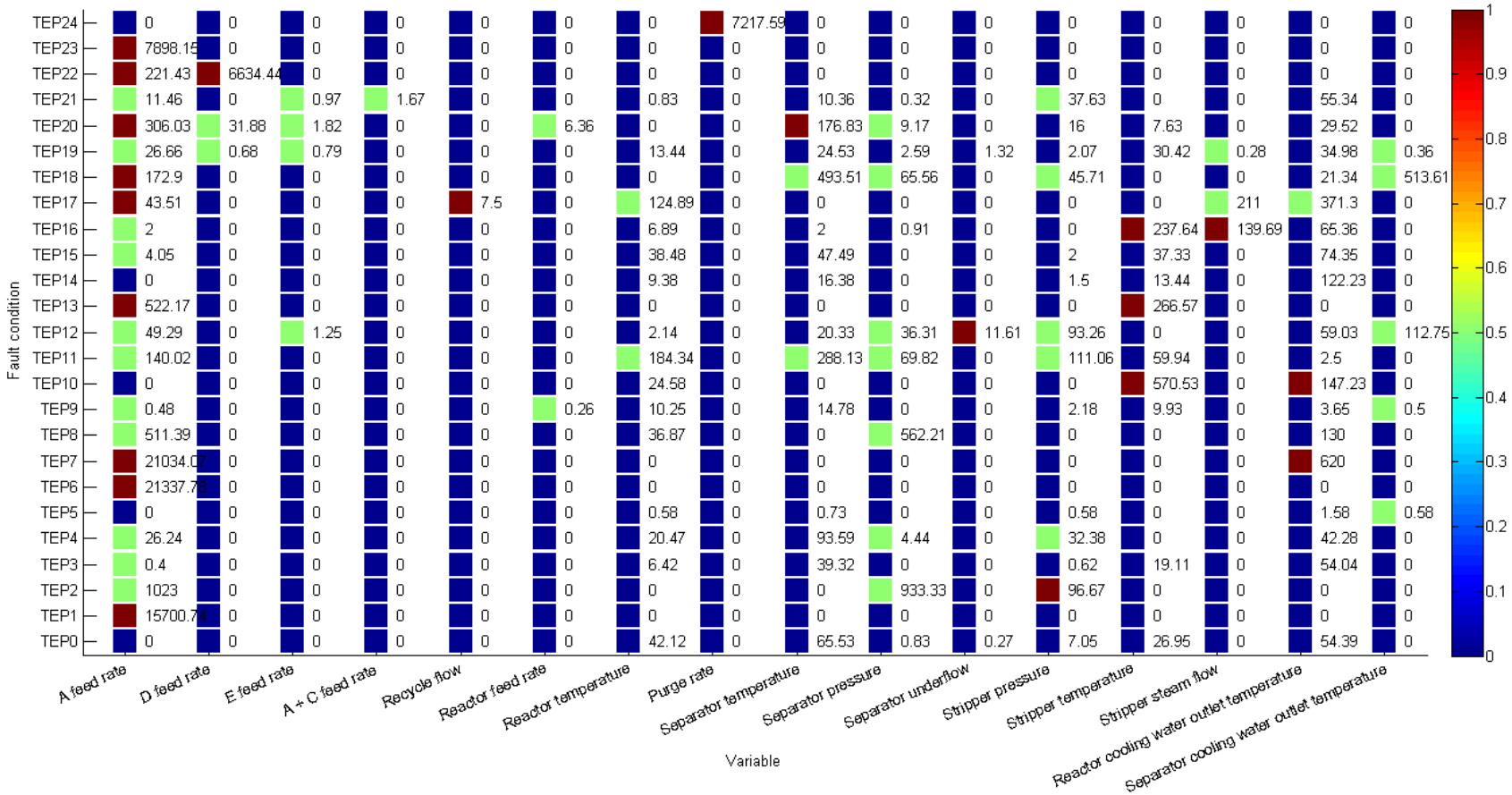

Figure 52: **Tennessee Eastman process: trees-with-bagging variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

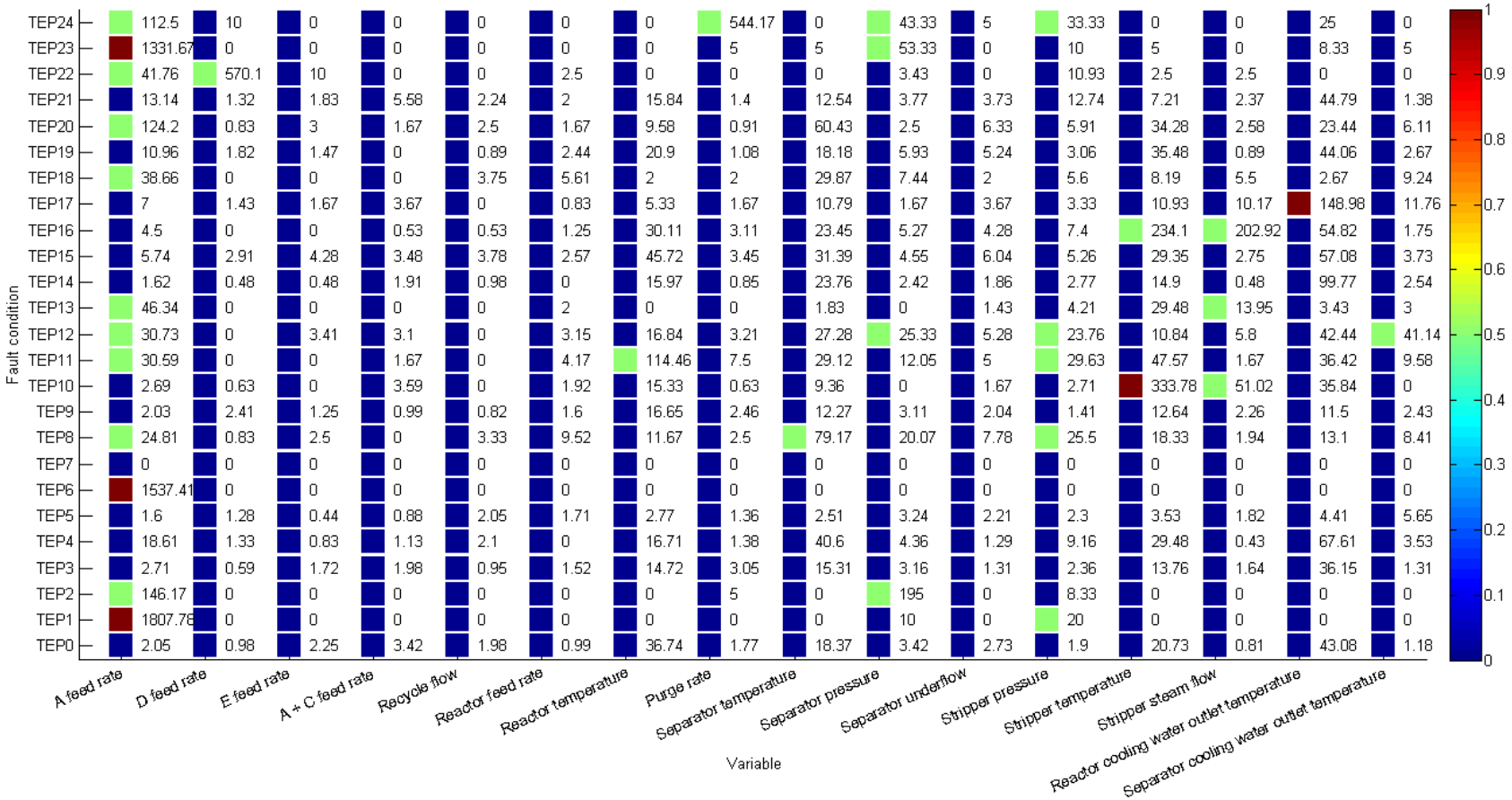

Figure 53: **Tennessee Eastman process: random forests variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

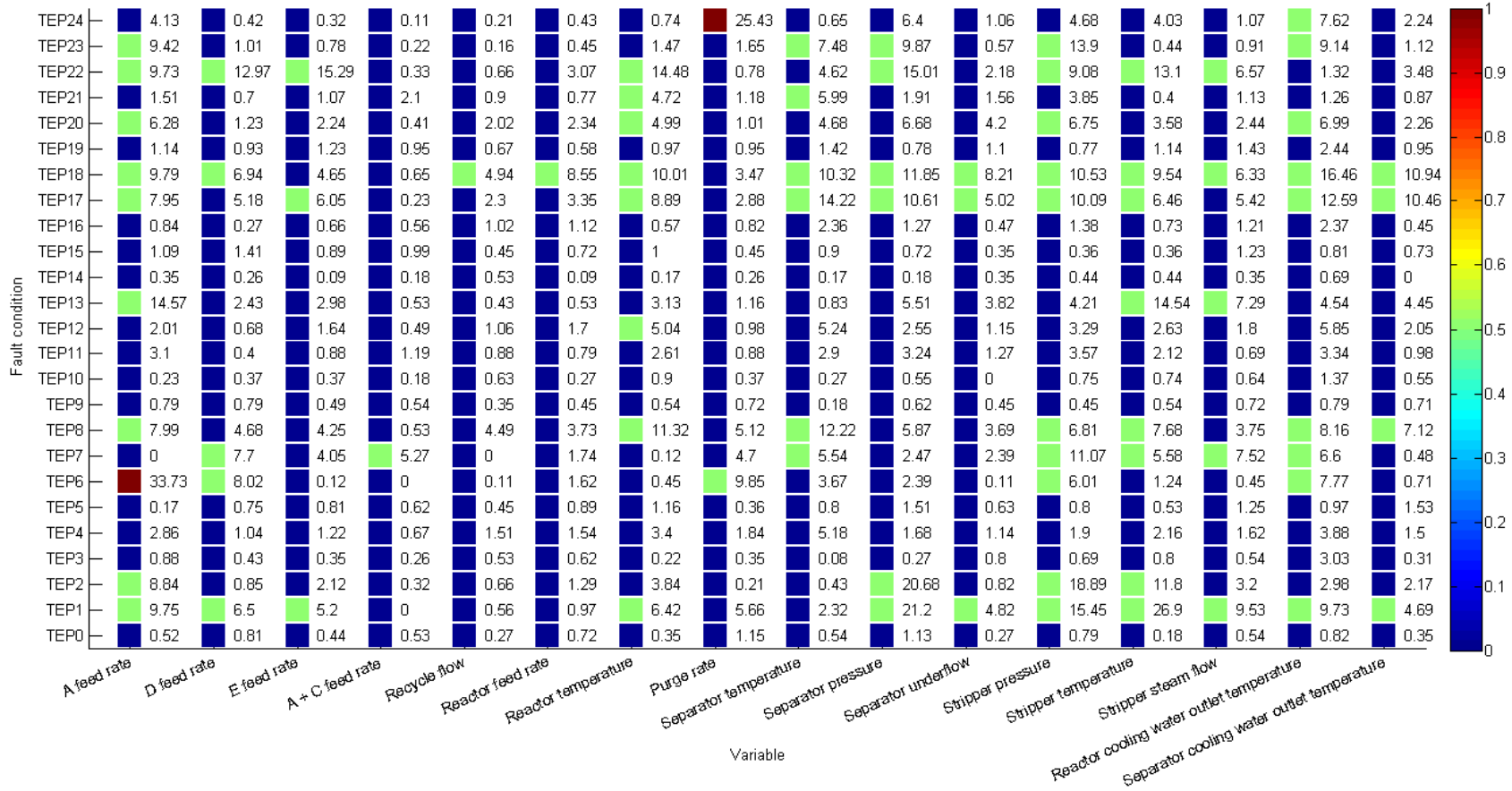

Figure 54: **Tennessee Eastman process: ELM-with-bagging at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

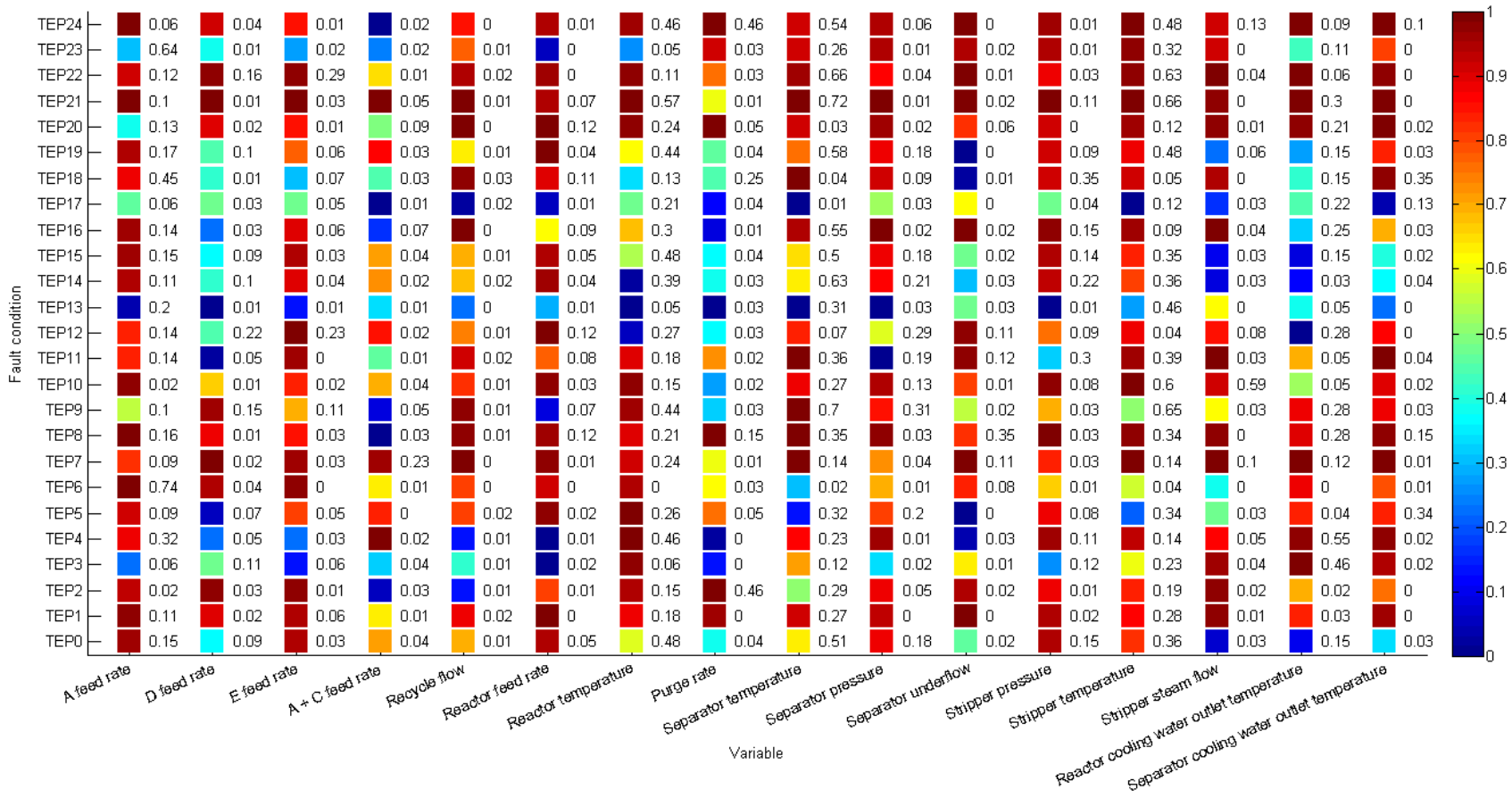

Figure 55: **Tennessee Eastman process: CVA biplot variable importance at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

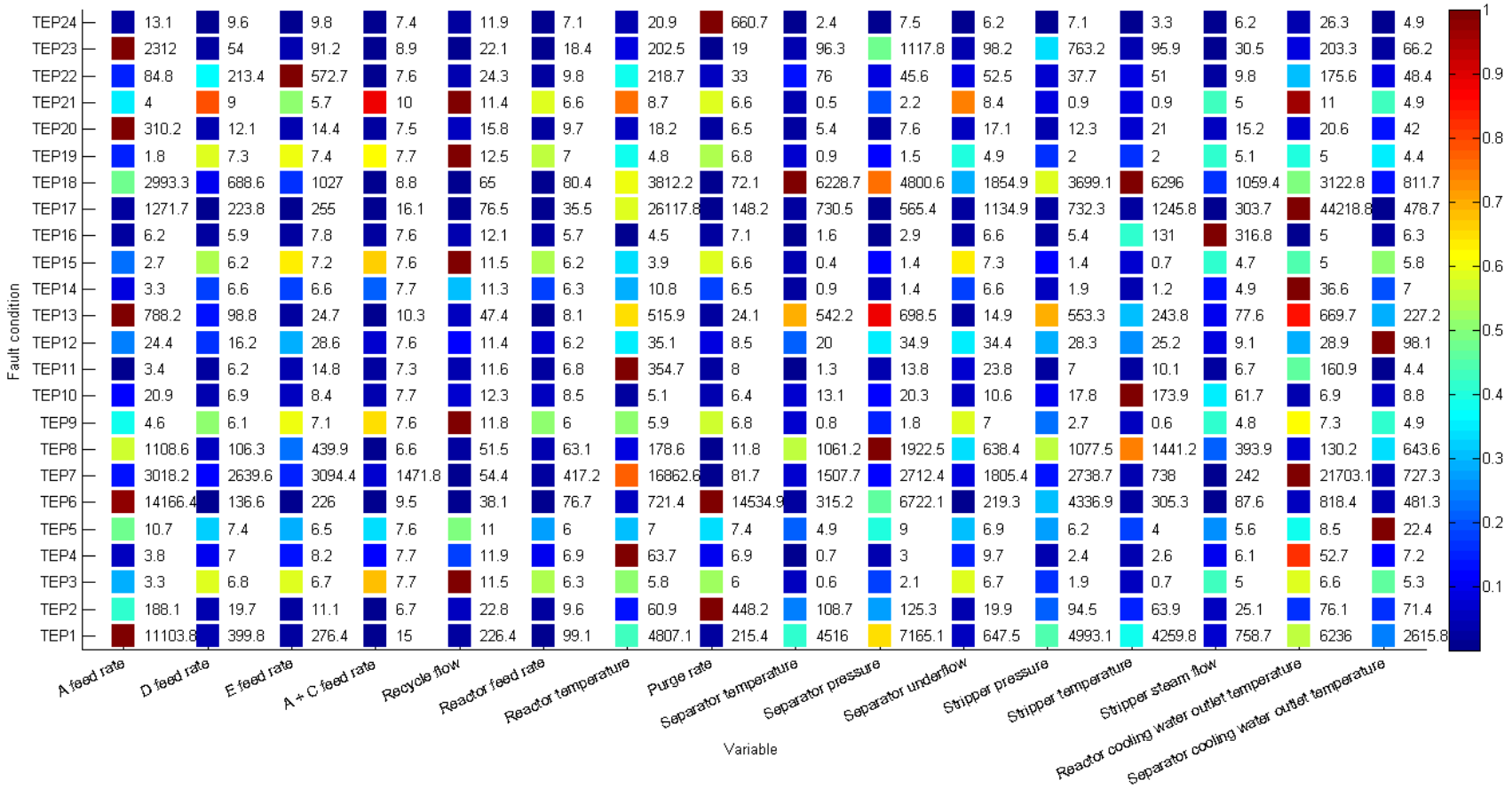

Figure 56: **Tennessee Eastman process: PCA SPE variable contributions at a confidence level of 0.99**

Stellenbosch University http://scholar.sun.ac.za

5.3.4 Summary

From the variable importance technique evaluation it is evident that there is no single variable importance technique that is effective in identifying the important variables for all potential fault conditions. It was found that all the techniques performed very similarly, with the linear discriminant analysis variable importance technique possibly performing slightly worse that the other techniques when considering the results of all the case studies. For the CVA biplot results, a fair bit of visual interpretation of the individual variable CVA biplots were also required in addition to the axes predictivities and adequacy values for the technique to be effectively used as a variable importance measure. This visual inspection of the results did, however, prove invaluable in gaining an understanding of the important variables for each fault condition. Due to the nature of the data being analysed, different techniques may be required considering the assumptions the techniques are based upon:

- For the simple multivariate time series data, being independent and identically distributed with a normal distribution, the trees-with-bagging and random forests variable importance analysis techniques performed the best of the techniques tested.
- For the simple multivariate process data, not being independent and identically distributed with a normal distribution, the trees-with-bagging variable importance analysis technique again performed the best of the techniques tested.
- Furthermore, although the data characteristics of the data set violated the assumptions of the trees-with-bagging variable importance analysis technique, the technique proved to be exceptionally robust in this regard and was still able to correctly identify the majority of the important variables.
- For the Tennessee Eastman process data, having a mixture of data characteristics with regards to being independent and identically distributed and having a normal distribution, including high levels of collinearity, very similar performance was achieved by all the variable importance techniques evaluated.
- As before, although the data characteristics of the data set violated the assumptions of all of the variable importance analysis techniques, the techniques proved to be fairly robust in this regard and were still able to correctly identify a large portion of the important variables.

As with the fault detection techniques, it is suggested that multiple variable importance techniques be run in parallel and their results be interpreted in conjunction with expert process knowledge.

Following the success of the trees-with-bagging variable importance analysis approach, it was proposed to replace the CART algorithm with the ELM algorithm for classification. Although it has been shown that neither neural network nor CART models have a clear advantage of one over the other when considering prediction accuracy (Razi and Athappilly, 2005), neural network methods do have the advantage of good generalization. As expected, the proposed ELM-based variable importance analysis technique was found to be a viable alternative to the CART based trees-with-bagging variable importance analysis technique, with both techniques performing very similarly for all case studies evaluated.

Stellenbosch University http://scholar.sun.ac.za

As expected, correlation in the data seemed to be a challenge for all the variable importance techniques with the techniques not necessarily being able to always distinguish between association and causation. The importance of monitoring as many of the process variables as possible was also highlighted as it is impossible to identify truly important variables if they are not being monitored. Monitoring many variables simultaneously does, however, have a drawback to it. Not only does it become increasingly difficult to identify the correct variable where the fault condition originates from, but visualisation also becomes more complex with the CVA biplot representations becoming very crowded and difficult to interpret. From the Tennessee Eastman process case study results it was noted that an absolute "correct" answer isn't always obtainable regarding variable importance. NOC data was first used to calibrate the techniques for the various variables. This was followed by the calculation of the performance statistics for the various fault conditions, with a "feel" being obtained for the magnitude of the performance statistics. This in turn was used to determine heuristic rules which were used to determine whether or not a specific variable importance technique could successfully identify the important variables for a specific fault condition.

When comparing the variable importance analysis results to the PCA SPE variable contribution results, it is evident that the variable importance analysis techniques are at least as capable at identifying important variables as the PCA SPE variable contribution techniques are, often being able to identify the important variables where the PCA SPE variable contributions could not. From this is concluded that the variable importance analysis techniques are a viable alternative to the PCA SPE variable contribution techniques. The advantage of the variable importance analysis techniques over the PCA SPE variable contribution techniques lies with the fact that they can be used not only for analysing potentially interesting process events detected by change point detection, when statistical data-based fault detection is unsuitable, but also when the statistical data-based fault detection models are inaccurate or the contribution plots unreliable. As with the fault detection techniques, it is again suggested that for complex processes having many variables requiring monitoring, the complexity of the problem be reduced through the development of process causality maps.

Stellenbosch University http://scholar.sun.ac.za

6 Analytical methodology and concentrator causality maps

Although various methodologies and techniques exist for multivariate process performance monitoring, no universal analytical procedure is available to "solve all problems". Typically different techniques are used by different users based on their knowledge or lack of understanding of the techniques and the availability or complexity of the data analysis software. The aim is, therefore, to define a generic analytical methodology for multivariate process performance monitoring, with application to the mineral processing industry, to ensure that the correct techniques and appropriate interpretations are always performed at the correct stage of any such analysis.

More specifically, this chapter deals with the definition of a multivariate process performance monitoring methodology based on the integration of process causality maps, statistical data-based fault detection, change point detection and variable importance analysis.

6.1 Analytical methodology

With many events occurring simultaneously within any one time frame in a process, the idea is not to try and focus on all events simultaneously in isolation but to rather try and understand how different events affect each other and thus identify the root cause of a problem. Probably as important as this is the requirement for constructing a view of events showing the interaction between different variables and process operations: not just identifying the root cause of a problem but also showing the cause and effect relationships throughout the process. It is often easier to encourage people to action by not only showing them what to fix, but also why it is important.

Furthermore, when dealing with large data sets it is important to determine where the useful information in the data can be found. Typically, should clusters form when visualising the data, the important information can usually be found in the transitions between the various clusters and not only within the clusters themselves. The reasons for these shifts in process operation could then be investigated through examination of the underlying multivariate process model. These process models can also be used to monitor either the current process values in order to determine if a fault has occurred or the predicted process values in order to determine if a fault is going to occur.

For the multivariate process performance monitoring of mineral processing plants the data analysis methodology below is proposed.

6.1.1 Problem identification

Probably the most important step to any data analysis methodology is the definition of the data analysis objective. Having a clearly and explicitly defined data analysis objective is especially important in ensuring that time is not wasted investigating unimportant interestingness in the data, but rather spent finding potential solutions to specific problems. To this end, process knowledge and operational experience and expertise are critical requirements for ensuring effective problem identification.

Stellenbosch University http://scholar.sun.ac.za

As part of the problem identification, the following should also be clearly defined:

- The detected event. This could be the result of the use of change point detection, statistical data-based fault detection (if NOC data exist) or even user observation (through the visual classification/clustering of data).
- The process to be monitored, including the analysis boundaries.
- The relevant process causality maps and associated process objectives.
- Any available process measurements relevant to the analysis.
- Key performance indicators relevant to the analysis.

6.1.2 Data collection

Following problem identification, data needs to be collected for all identified process measurements and KPIs. Data is typically collected from the following data sources: process historians (typically containing high frequency data), laboratory databases (typically containing low frequency data) and various manually populated log sheets or spreadsheets (typically containing low frequency data). For data collection, the collection period and sampling frequency need to be appropriately selected:

- The data collection period should include both normal/reference operating conditions (NOC) data as well as known and unknown event/abnormal/fault conditions data.
- The data sampling frequency should be as high as possible as data can always be down-sampled at a later stage.
- Since the developed model will only be applicable to the conditions under which the data are collected, the data sets should encompass the range from low to high values of the process variables yielding the desired outputs.

For some processes, selection of the traditionally defined NOC data is not always straightforward. Usually NOC data are defined as conditions where specific measurements are within a predetermined specification. If these conditions cannot be maintained for reasonable periods of time, or if the process is naturally unstable, it may be difficult to select a sufficiently large NOC data set. If this is the case, the NOC data may simply be defined as data associated with stable process operation, irrespective of whether specific measurements are within predetermined specifications.

6.1.3 Data pre-treatment

Following data collection, the data is subjected to data validation and characterisation. At this stage it is important to remember that the analysis results are only as good as the data on which it is based. If the data analysed is of poor quality, it can only be expected that the results would be unreliable. However, if the data is properly validated and cleaned prior to being analysed, it can be expected that the results would be reliable. Furthermore, different data analysis techniques make different assumptions regarding the data being analysed. It is therefore important to properly characterise the data prior to analysis, thus ensuring that the results can be interpreted in the correct context depending on the combination of data characteristics and analysis assumptions.

Stellenbosch University http://scholar.sun.ac.za

For data validation, gross error detection is initially performed on the data, marking the following data qualities as “bad”:

- Missing, not a number and infinite data – missing data.
- Process measurement values that are associated with equipment that is not running – not running data.
- Process measurement values exceeding the instrument measurement high or low detection limits – high/low data.
- Process measurement values that do not update over a predetermined time window – not updating data.
- Process measurement values where the rate of change between consecutive samples in a predetermine time window exceeds a predetermine threshold – rate of change exceeded data.
- Data failing other user defined outlier, mass balance, soft-sensing or data sensibility checks.

For data qualities marked as “bad”, the data is either removed from the data set, or replaced. If the data is removed, new data sets are defined, separated by the “bad” quality data. If the “bad” quality data is to be replaced, various strategies could be used: mean of data, linear model, PCA, neural network model, etc. For NOC data, no abnormal events should be present in the data set. Statistically this implies that no more than 5% of the NOC data should exceed the 2σ threshold. Should more than 5% of the data exceed this threshold following gross error detection, such data should be treated as abnormal event situations and removed from the data set (Nijhuis et al., 1997).

Depending on the underlying assumptions of the data analysis techniques used, technique specific data characterisation and pre-processing should be applied:

- The data could be auto-scaled to zero mean and unit variance.
- The data could be tested to determine if it is normally distributed.
- The data could be tested to determine if it is exponentially distributed.
- The data could be tested to determine if it is independent and identically distributed or auto-correlated. If the data is found to be auto-correlated, the results could also be used to determine sensible data aggregation intervals.
- The data could be tested to determine if it is collinear, highlighting the presence of cross-correlation in the data and raising a warning regarding the accuracy of subsequent results. Alternatively, for collinear data having similar autocorrelation profiles, collinearity could be removed through the PCA based transformation of the data.
- Using a combination of correlation coefficients between time-delayed variables and lag-plots, delays between inputs and outputs can be determined and the data adjusted accordingly.
- Noisy data could be filtered using e.g. zero-phase forward and reverse digital filters, PCA, etc. Unfortunately, filtering will remove information from the data and if valuable information is removed, the filter residuals should also be included in the data set to be analysed.

Stellenbosch University http://scholar.sun.ac.za

- If down-sampling of the data is required, data features such as mean, median, standard deviation, kurtosis, skewness, etc. could be calculated as an alternative to simply selecting equally spaced subsamples from the data set.
- Transformed data could also be prepared for use in the data analysis in the form of reconstructed data, principal components, model residuals, etc.

6.1.4 Data analysis

Given the problem identification, the detected event, the first step of data analysis is to visualise the event data through e.g. time series trends, normal distribution comparison plots, etc. This is important as it familiarises the analyst with the relevant event data and also defines the context of the event. If the event was detected through visual observation, specific events in the data need to be determined through the application of change point detection and confirmed through the use of techniques such as median significance analysis.

For complex systems, a holistic overview of the problem is first created. This overview is used to get a feel for the complexity of the problem as well as to identify potential focal points upfront. The analysis overview is obtained through the creation of the combined data set containing data from all the process measurements and KPIs relevant to the analysis and the subsequent application of change point detection, statistical data-based fault detection (if NOC data exist) and variable importance analysis. It should be noted that whenever data analysis techniques are applied to large data sets, the possibility also exists to either down-sample the data or to randomly sub-sample the data prior to data analysis and then compare data analysis results from different random subsets.

Following the process causality map hierarchy, starting with the main identified event:

1. Use the process causality map to determine which variables are relevant to the variable containing the event as well as the context of these variables within the process, creating a reduced data set. This will not only reduce the size of the data set to be analysed, but also allow for more robust, concise results.
2. Apply statistical data-based fault detection to the reduced data set, creating models that can be used for future detection and analysis of similar events.
3. Visualise the data using CVA biplots to determine if it is possible to visually distinguish between the different event classes and to identify potential important variables.
4. In support of CVA biplots, perform procedural variable importance analysis to identify important variables (root causes). Both between-class and within-class variable importance analysis could be performed. This is especially important as it allows the one to one analysis of data classes for determining important variables between only 2 data classes as opposed to 3 or more when using CVA biplots.
5. Apply median significance and one-way ANOVA analysis to the identified important variables in support of the variable importance analysis findings.
6. Apply change point detection to the important variables in order to determine if similar change points can be detected to those in the variable containing the event. Similarly, it would be

Stellenbosch University http://scholar.sun.ac.za

interesting to see if the variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others. Change point detection could be applied to individual time series data, multivariate data sets or features of either individual time series data or multivariate data sets.

7. Given the context of the identified important variables within the process causality map hierarchy, repeat from step 1 until the root cause has been identified.

6.1.5 Implementation of findings

Depending on the outcome of the data analysis, various alternatives exist with regards to implementation of the findings. First and foremost, where possible, a solution to the identified root cause of a problem should be found. Focus should be on finding the simplest, easiest workable solution. To this end a detailed implementation plan should be developed and corrective action implemented accordingly. Following this, the process should be monitored to ensure no repeat of the problem through continued success of the corrective action. If required, new data could be evaluating using the developed reference models. New abnormal process conditions (events) and underperformance could then be identified and the analysis repeated.

One of the often overlooked, but potentially extremely valuable outcomes is the opportunity to learn from the aforementioned data analysis. Not only could the data analysis be used to identify the root cause of a problem, but it could also be used to gain insight into the cause and effect relationship within the process and highlight previously unknown relationships within the process.

6.2 Concentrator case study

The mineral processing plant under investigation concentrates UG2 ore containing platinum group metals. The process consists of three grinding circuits with their associated flotation circuits (Figure 57). The primary grinding circuit consists of a closed circuit fully autogenous primary mill with the mill product classified by a vibrating screen. The oversize/coarse particles from the vibrating screen are recycled back to the primary mill for regrinding with the undersize/fine particles reporting to the chrome cyclone. The underflow from the chrome cyclone reports to an open circuit chrome ball mill (secondary grinding circuit) with the mill product reporting to the flash flotation circuit. The overflow from the chrome cyclone reports to two closed circuit silica ball mills operating in parallel (tertiary grinding circuit) via the primary flotation circuit with the circuit product reporting to the scavenger flotation circuit.

The process is well instrumented with power measurements available on all mills, load measurements available on the primary and secondary mills and flow and density measurements available on all main process streams. Other miscellaneous measurements are also available with data logged at a high sampling frequency.

For this case study, the objective is to reliably detect a shift in the performance of the process and to accurately find the root cause of the shift. The shift to be detected is a higher than normal final tailings grade value for the process. The data required for the analysis are selected based on this problem

Stellenbosch University http://scholar.sun.ac.za

definition and data availability. The data used for the analyses originate from one of two sources, viz. laboratory or online instrumentation, which are available at significantly different sampling frequencies. Analytical results from the laboratory, such as tailings grade data or grinding circuit particle size distribution (PSD) data are only available at a rate of one sample per day (low frequency). In contrast to this, online process data, such as mill power measurements or slurry flow measurements are available at a rate of one sample every 10 seconds (high frequency). The historical data collected for the analysis spanned approximately 5 months of almost continuous process operation and capture a reasonable amount of process variability.

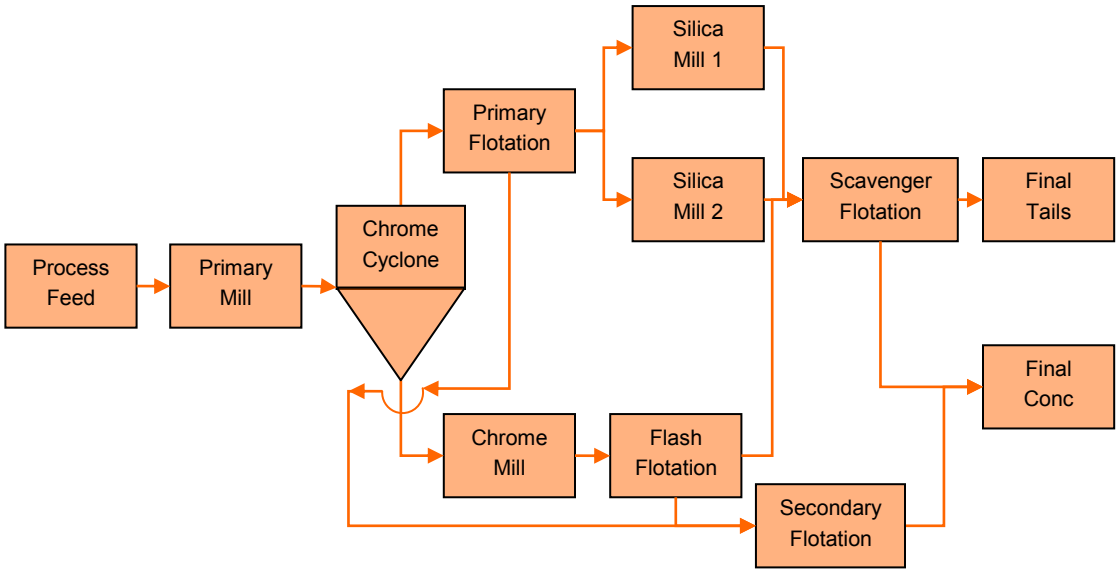

Figure 57: **Process flow diagram of a concentration plant at Anglo American Platinum**

6.3 Concentrator process causality maps

6.3.1 Crusher process causality map

The primary objective of a crusher is to reduce the particle size of the feed ore. Crushers operate on a repetitive cycle of classification and breakage events and are typically used in combination with screens to ensure (1) maximum size reduction, (2) maximum product of a specified size, (3) top size reduction or (4) top size control with minimum fines production (Napier-Munn et al., 1999).

Crushers most commonly used on platinum concentrating processes include jaw crushers, cone crushers and roll crushers (especially the High Pressure Grinding Rolls). Whereas jaw crushers are limited in terms of capacity and size reduction due to the swell factor of broken particles, cone crushers provide a solution to this by wrapping the jaw crusher profile around a central axis. This allows increased cross sectional area at finer gaps as well as opportunities for multiple impacts. By choke feeding a cone crusher additional benefits are to be gained from using particles to break each other instead of abrading

Stellenbosch University http://scholar.sun.ac.za

expensive alloy steel, decreasing the product size and increasing the severity of breakage. Choke feeding is normally optimised through good feed rate control and a level sensor in the crusher mouth to detect when maximum flow through is exceeded.

For crushers the following are important machine dimensions (Napier-Munn et al., 1999):

- Throat dimension
- Open-side setting (largest fall through aperture)
- Closed-side setting (minimum opening)
- Liner dimensions
- Crusher gap (for cone crushers)
- Liner profiles (for cone crushers)

Other important process variables include throughput (the most critical operational variable), crusher level (required for crusher choke feed optimisation), feed PSD, power draw and product PSD. Of these the most important production KPIs are:

- Crusher gap
- Power draw

In general for cone crushers the product PSD (primary objective) is dominated by the closed-side setting, but also influenced by throughput and feed PSD. Power draw gives an indication of the energy required to reduce the ore to the required PSD and is predominantly affected by throughput and feed PSD. At the highest level the crusher product PSD is therefore a function of the power draw and crusher closed-side setting. In turn, power draw is primarily a function of throughput and feed PSD but can also be affected by a change in the hardness of the feed or ore something being wrong with the crusher (such as liner wear). Liner wear affects the maximum power draw and productivity which the crusher can attain and can be accounted for by considering the hours of operation and liner characteristics. Lastly, power draw is also influenced by the liner profile with the liner profile further influencing throughput and product PSD. Throughput, being a sub function of power draw, is therefore a function of the liner profile, the crusher level and the crusher gap. From this we can now derive a generic process causality map for crushing (Figure 58).

Stellenbosch University http://scholar.sun.ac.za

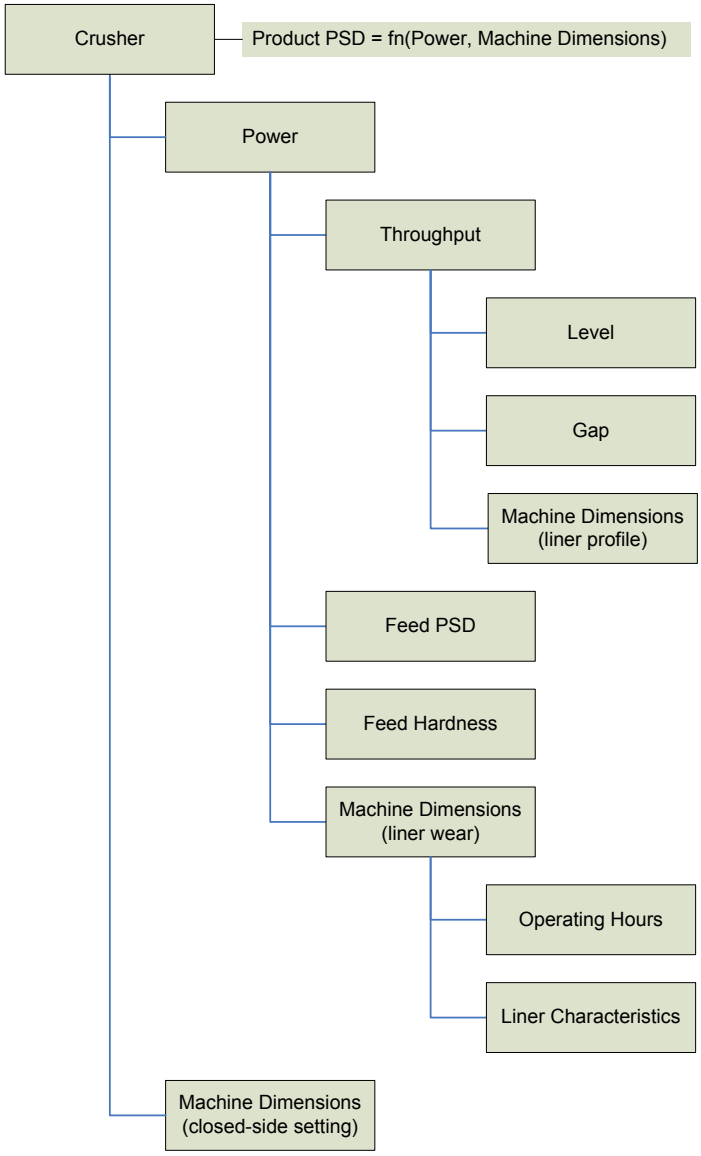

Figure 58: **Generic crusher process causality map**

6.3.2 Milling process causality map

The primary objective of a mill is to reduce the size of the feed material with the aim of producing as fine a grind as economically viable. The design criteria of the mill typically define what minimum fineness of grind should be achieved. This design criterion should include the following important machine dimensions (Napier-Munn et al., 1999):

- Aspect ratio (diameter to length ratio)
- Method by which product is discharged
- Shape and material composition of lifters and liners
- Liner wear

Stellenbosch University http://scholar.sun.ac.za

In a fully autogenous mill, the feed PSD and ore hardness dictate the volume and size distribution of the grinding media. The grinding process having essentially three components:

- Collision frequency (breakage rate)
- Ore PSD after collision (appearance distribution function)
- Particle transport out of the mill (discharge rate)

It is important to note that material transport is affected by the mill discharge grate. It has been found that higher open areas result in higher flow rates for the same slurry level in the mill. Also, holes at a larger radial distance from the centre of rotation of the mill will pass higher flows at a constant slurry level. In practice, however, depending on the geometry and open area of the grate, the slurry level in the mill adjusts itself given a set flow rate of water and solids into the mill. Smaller particles (<2.5-5 mm) are also increasingly influenced by the flow of water through the mill resulting in the probability that the particles will be broken being decreased as their size reduces.

In ball mills the mass of the balls dominates both the power draw and the grinding performance of the mill with the mill power draw remaining relatively constant and only reducing gradually over time as the ball charge slowly wears down.

For a mill at equilibrium, changes in feed rate, feed PSD or feed hardness will cause disturbances in its behaviour and need to be actively controlled. However, of these variables feed PSD and ore hardness are sometimes uncontrolled and these uncontrolled variations are consistently found to negatively impact milling performance. Changes in feed rate affect mill charge volume and power draw, these variables being closely linked. Under normal operating conditions the following relationships exist between these variables:

- An increase in feed rate will lead to a rise in the mill charge volume, causing the mill power draw to increase.
- An increase in ore hardness, resulting in the ore not breaking down as rapidly as normal, will also lead to a rise in the mill charge volume, causing the mill power draw to increase.

The mill power draw therefore does not change in direct response to the feed rate, but rather due to changes in the mill charge volume.

In general, the performance of fully autogenous mills is better with a coarser feed, however, if large amounts of middle size material (25-50 mm) are present in the mill feed, mill performance will be poor. Any changes in the mill feed PSD will result in changes in the grinding media PSD which in turn will affect the mill breakage characteristics. Ore hardness has been found to be closely related to ore PSD with harder ore typically producing a coarser feed PSD to the mill. When the feed ore is either too soft or too hard, mill throughput may be limited.

The speed at which a mill turns directly influences ore breakage in the mill. An increase in mill speed will at first increase the amount of lift imparted to the charge, up to a maximum, after which it will reduce until

Stellenbosch University http://scholar.sun.ac.za

the lift is effectively zero (when the mill centrifuges). Operationally, impact breakage is increased as the mill speed is increased, resulting in increased coarse rock breakage. This in turn leads to higher throughputs. However, increased coarse rock breakage results in less ore grinding media which negatively influences fine grinding conditions, resulting in a coarser mill product and increased liner wear when rock on shell breakage occurs (resulting in increased grinding costs). Typically slurry hold-up in the mill also changes with mill speed, increasing as mill speed is increased.

Slurry hold-up in a mill can also be increased by closing the milling circuit with a fine classifier, returning the underflow from the classifier back to the mill. This typically leads to increased attrition grinding, decreased abrasion breakage and an increase in charge density which in turn will result in a slightly increased power draw and finer grind. However, too much slurry recycle can lead to slurry building up in a pool at the toe of the mill charge, leading to reduced impact breakage at the charge toe which will result in a drop in the mill power draw. This will lead to coarser ore building up inside the mill, the effect of which can only be corrected by reducing the feed rate to the mill and increasing the classifier cut point (or diverting some of the recycled material).

Milling performance is also commonly measured using a classic mill operating curve (Figure 59). In theory the mill operates along the mill operating curve – moving up and down the curve based on the relationship between the mill power and the mill load. In practice, however, many of these operating curves exist for a single mill based on the operating conditions of the mill at any point in time. The mill therefore not only operates along the mill operating curve but also moves around between the various mill operating curves.

Given these important mill operational relationships, Napier-Munn et al. (1999) lists the minimum recommended instrumentation/measurements for the effective control/optimisation of a fully autogenous mill as:

- Feed belt weightometer – to ensure that the feed rate can be maintained at a given level with minimum surging.
- Accurate mill power measurement – to ensure that the mill is operated close to maximum power of the motor, maximising throughput while minimising specific energy.
- Charge level indicator such as load cells or bearing pressure meters – to maintain the mill load at a level below the maximum volume of charge that can be accommodated in the mill.
- Discharge slurry density gauge – to maintain a constant discharge slurry rheology

Steel consumption – liner wear for fully autogenous mills and steel ball consumption for ball and semi-autogenous mills should be kept to a minimum. Low charge levels should specifically be avoided to limit the collision of balls with exposed lifters and liners.

Stellenbosch University http://scholar.sun.ac.za

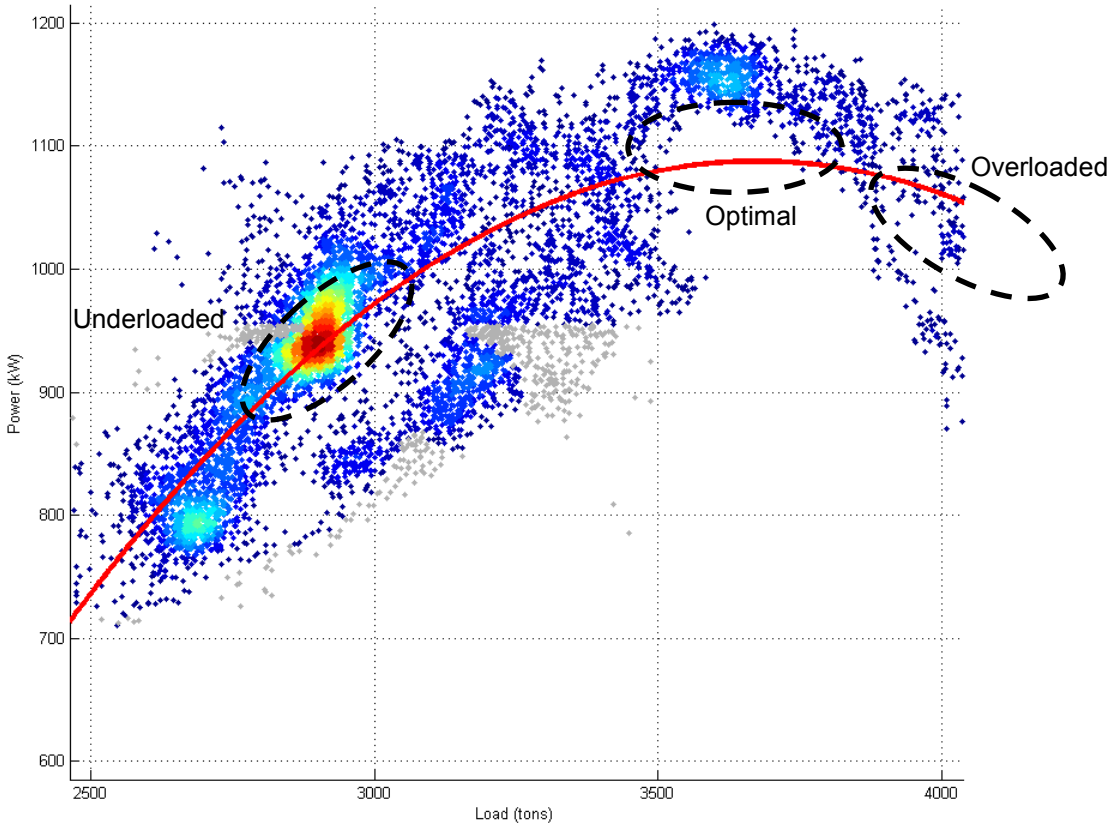

Figure 59: **Mill operating curve with key operational regions shown**

In addition to this it should be noted that the most important production KPIs for a mill are:

- Mill kWh/t -75 µm
  - Mill feed rate (tph) & stability (number of stops)
  - Mill power draw (kW) & stability (number of stops)
  - Grind (% -75 µm)
- In mill density (t/m^3)

In general for grinding mills the product PSD (primary objective) is dominated by changes in the throughput, feed PSD and feed hardness. Throughput also has a direct effect on mill power draw, mill load and, when combined with the inlet water ratio, in-mill density. Under normal operating conditions and increase in the mill throughput will lead to an increase in the mill load and an increase in the mill power draw. Similarly, an increase in ore hardness will also lead to an increase in the mill load and an increase in the mill power draw. At the highest level the mill product PSD is therefore a function of the power draw, the load and the in-mill density. All these sub functions are closely related and, in turn, a function of throughput, inlet water ratio, speed, re-circulating load, grinding media and the mill dimensions

Stellenbosch University http://scholar.sun.ac.za

(such as lifter profile). Whereas an increase in mill speed will typically lead to an increase in slurry hold-up, resulting in a finer grind, any changes in the mill grinding media PSD will affect the mill breakage characteristics which in turn will also affect the mill product PSD. Furthermore, for a fully autogenous mill the mill grinding media sub function is a function of ore hardness and the feed PSD (coarse/fine ore ratio). From this we can now derive a generic process causality map for milling (Figure 60).

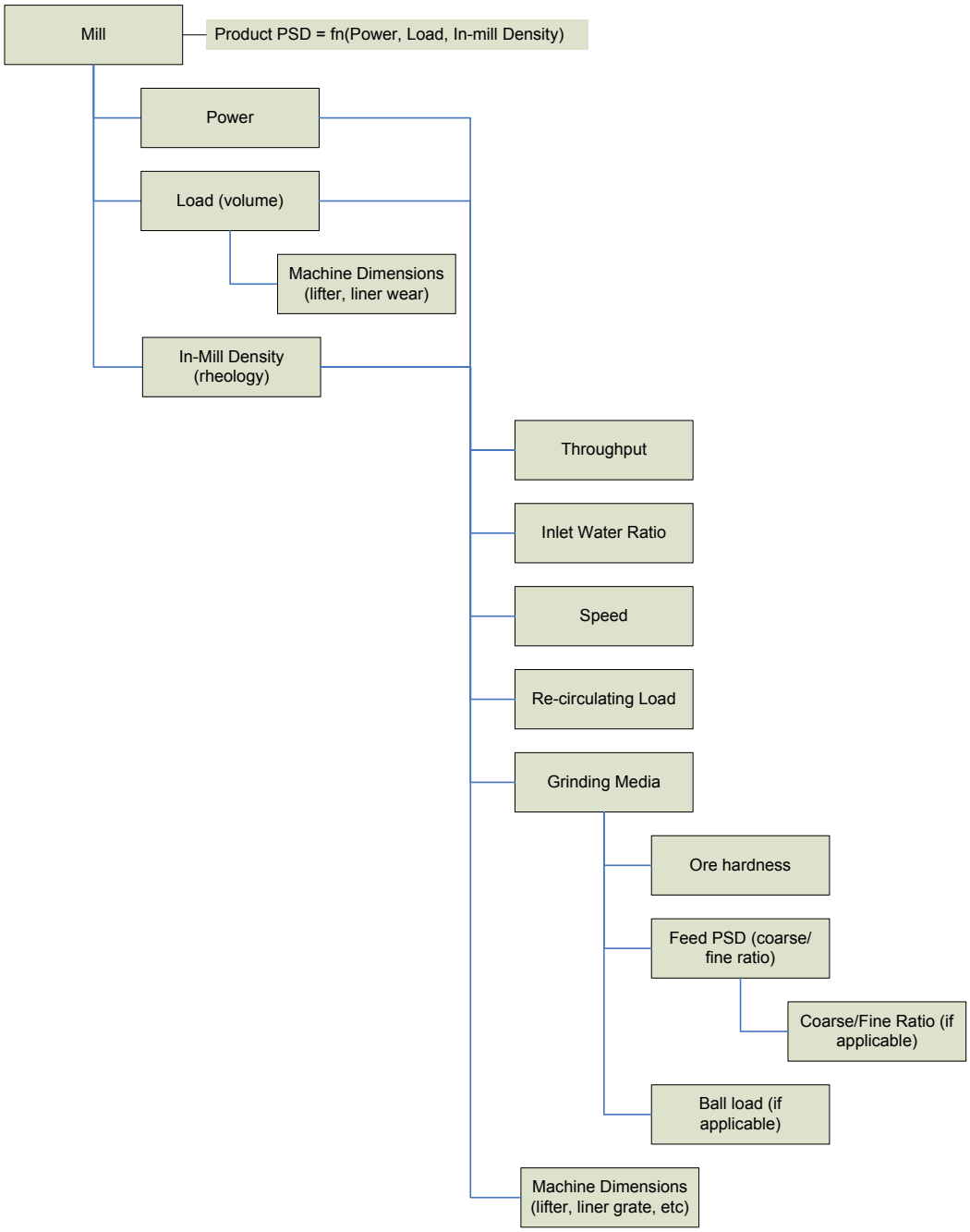

Figure 60: **Generic mill process causality map**

Stellenbosch University http://scholar.sun.ac.za

6.3.3 Cyclone process causality map

The primary objective of a cyclone is to classify its feed material into an overflow stream containing the finer/lighter material and an underflow stream containing the coarser/heavier material. A cyclone can also be used to densify (removing excess water) its feed material producing an overflow stream containing the excess water and an underflow stream containing the densified slurry.

Cyclones are typically used in a closed circuit where the circuit product must meet some particle size criterion before it is allowed to continue downstream. In an open circuit, cyclones are used to ensure appropriate overflow and underflow stream characteristics for further processing. For optimisation, cyclones are usually the single most effective instrument for tuning circuit performance (Napier-Munn et al., 1999). Substantial improvements in performance can typically be gained through relatively easy and cheap changes in the classification process such as modifying the water content of the cyclone feed, changing the cyclone apex (spigot) size or changing the vortex finder size. These variables are defined in the main classes of variables available for optimising cyclone performance namely cyclone geometry and cyclone feed conditions.

Cyclone efficiency is characterised by the proportion of a given size of solids which reports to the underflow or overflow products. This measure can be reduced to achieve a desired, stable, mass split and cut-point. The mass split simply defines the desired amount of solids in the overflow and underflow respectively whereas the cut-size (separation size) is normally reported as d50c, defining the particle size of the material which divides equally between overflow and underflow due to classification forces only. An increase in cyclone cut-size can result from any of the following: reducing the size of the apex or increasing the size of the vortex finder; for larger, low pressure cyclones, inclining the cyclone to the vertical; lowering the flow rate or pressure to the cyclone appropriately; or increasing the solids concentration of the cyclone feed. Since this solids-classification behaviour is strongly correlated with the total volume and water split achieved by the cyclone, many of these factors also influence the proportion of water reporting to the underflow of the cyclone. The proportion of water reporting to the underflow of the cyclone can be reduced by installing a smaller apex or larger vortex finder, by increasing the pressure of the cyclone or for larger cyclones by inclining the cyclone. Increasing the solids concentration of the feed to the cyclone will tend to increase the portion of slurry reporting to the cyclone underflow.

For performance monitoring of a cyclone, visual assessment is confined to observing the characteristics of the underflow discharge. Instrumentally, feed flow rate, density and pressure drop are typically available. However, the performance of the circuit which the cyclone is used to control is often measured as throughput and/or final product size distribution – none of which is usually available online. Reasons for cyclone underperformance include:

- Leakage of coarse material across the top of the cyclone to the overflow
- Trapping of or bypass of fines, causing fines to report to the underflow
- Inappropriate operating conditions such as excessive feed viscosity, roping or worn cyclone parts (particularly the apex and vortex finder)

Stellenbosch University http://scholar.sun.ac.za

Given these important cyclone operational relationships, Napier-Munn et al. (1999) list the following options as the most common with which to manipulate the cyclone performance:

- Number of cyclones – opening an additional cyclone in a cyclone nest will reduce the effective flow rate and pressure per cyclone, increasing the d50c and increasing the water recovery to the cyclone underflow.
- Apex – reducing the size of the apex will reduce the water recovery to the underflow of the cyclone and increase the d50c.
- Vortex finder – increasing the size of the vortex finder will reduce the water recovery to the underflow of the cyclone and increase the d50c.
- Feed density – increasing the density of the cyclone feed will increase the volume flow rate to the cyclone underflow, increasing the d50c, and may improve cyclone efficiency.
- Cyclone diameter – increasing the diameter of the cyclone, will increase the d50c and generally reduce cyclone wear and pumping costs.
- Cyclone inclination – inclining larger cyclones to 45^{o} or more to the vertical will increase the d50c and reduce water recovery to the cyclone underflow.

In addition to this it should be noted that the most important production KPIs for a cyclone are:

- Relative feed density (t/m^3)
- Feed flow rate (tph)
- Pressure (kPa)

In general for cyclones the mass split and cut-point (primary objective) is dominated by changes in the cyclone geometry and feed conditions. A strong relationship also exists between the operating pressure of a cyclone and the cyclone feed flow rate and geometry with an increase in cyclone cut-size being achieved by reducing the size of the apex, increasing the size of the vortex finder or lowering the cyclone feed flow rate. At the highest level the cyclone primary objective is therefore a function of the operating pressure and the feed PSD with the operating pressure, in turn, being a function of the cyclone feed flow rate, feed density and cyclone geometry. From this we can now derive a generic process causality map for a cyclone (Figure 61).

Stellenbosch University http://scholar.sun.ac.za

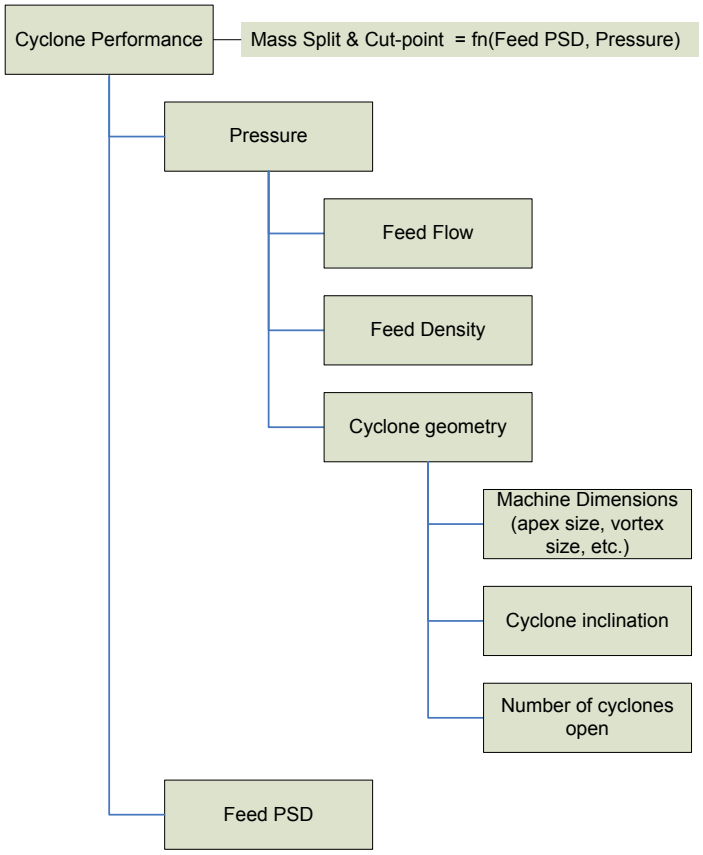

Figure 61: **Generic cyclone process causality map**

6.3.4 Flotation process causality map

The primary objective of froth flotation is to separate the mineral rich material (concentrate) from the gangue material (tails). Froth flotation is a physicochemical method of concentrating ground ores based on the different hydrophobicity of the particles to be separated. Since hydrophobicity of minerals is rarely naturally occurring the process involves the chemical treatment of an ore pulp to create conditions favourable for the attachment of pre-determined mineral particles to air bubbles – the elementary step of particle capture by a rising bubble being considered by some as the most important sub-process contributing to flotation performance. The air bubbles subsequently carry the selected minerals to the surface of the pulp, there forming a stabilised froth (concentrate) which is skimmed off.

For flotation, the effect of particle size on flotation recovery is significant – the particle size to be floated should be rather small so that the particles can be levitated by the bubbles produced. It has been shown that there exists a certain size range, varying with the mineral properties such as density and liberation, in which optimum results may be obtained in mineral processing. This range is in the order of 10-100 µm (Trahar and Warren, 1976). Due to this, flotation operates quite effectively with finely pulverized ores, making it an important operation in the treatment of low-grade or complex ores where, owing to liberation of grains, size reduction to finer particle sizes is compulsory.

Stellenbosch University http://scholar.sun.ac.za

For the required product quality to be obtained from flotation important factors include (Matis et al., 1993):

- Proper flotation cell design
- Appropriate operation of the conditioning stage
- Good selection of chemical reagents – usually classified as collectors, modifiers (activators, depressants, dispersants, pH regulators, coagulants, etc.) and frothers
    - Collectors promote or enhance the PGM particles ability to stick to the bubbles
    - Activators assist or work with collectors to enhance the collection of minerals to the froth, activating the particle surface and making it more receptive to a collector
    - Depressants increase the hydrophilic property of gangue material thus “depressing” the collection of this material to the froth
    - Promotors are typically used as secondary collectors, enhancing the valuable particles ability to float
    - Frothers stabilise the bubbles in the froth layer allowing more particles to attach themselves to the bubble before it bursts

Since phenomena that occur in the froth phase are known to significantly affect the results of the flotation process, important operational parameters include: air flow rate, frother dosage, air-bubble size distribution and mineral size and concentration in pulp. Furthermore, impeller speed, air rate and particle size also influences the concentration of gangue in the uppermost pulp layers – rotational speed and thus intensity of turbulence having a strong influence on the concentrate.

Hadler et al. (2010) showed that mass pull, a key performance indicator in flotation performance defined as the proportion of the feed material reporting to the concentrate, is affected by changes in froth structure and stability which in turn are affected by changes in operating parameters such as air flow rate and froth depth. It has been shown that optimising froth stability through variations in flotation cell aeration, flotation performance can be improved (Figure 62), whether in terms of concentrate grade, mineral recovery, or both (Hadler et al., 2006).

Stellenbosch University http://scholar.sun.ac.za

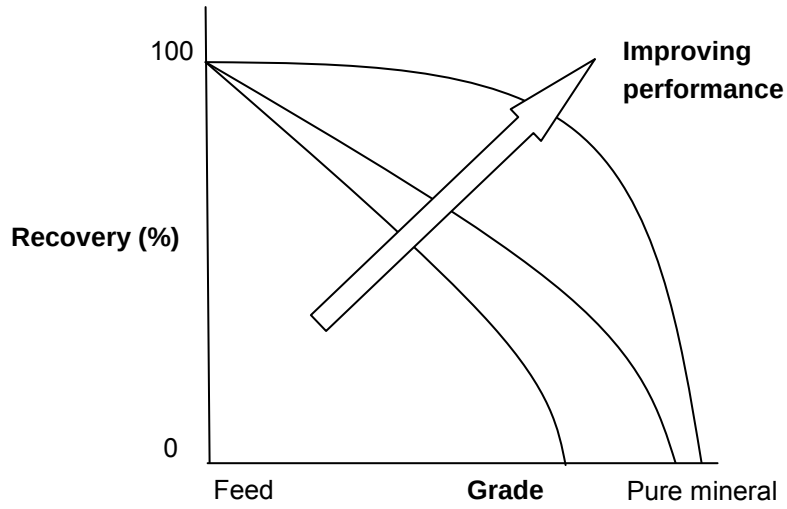

Figure 62: **Froth flotation grade/recovery curves showing performance improvement**

Froth stability is usually quantified using air recovery: the fraction of air entering a flotation cell that overflows the cell lip in un-burst bubbles. A higher air recovery typically indicates a more stable froth. Furthermore, low air flow rates produce froths that are highly laden and well drained but with low mobility. This results in bubbles that largely collapse before overflowing the cell lip, yielding low air and mineral recoveries but high concentrate grades. In contrast to this, high air flow rates produce high water content froths that are free flowing and unstable due to low bubble loadings causing bubbles to burst before overflowing the cell lip. As with low air flow rates, high air flow rates also lead to low air recoveries, however, reasonable mineral recoveries and low concentrate grades are obtained at high air flow rates. At intermediate air flow rates, a balanced, stable froth can be obtained. This yields a peak in air recovery, a high mineral recovery and only a somewhat compromised concentrate grade. Understanding the changes in froth stability with variations in operating conditions such as air flow rate can therefore lead to improvements in process performance (Hadler et al., 2010).

In addition to this it should be noted that the most important production KPIs for flotation are:

- Feed
  - Relative feed density (t/m^3)
  - Feed flow rate (tph)
- Mass pull
  - Froth depth (mm)
  - Air flow rate (m^3/hr)
- Reagents

In general for flotation the recovery and grade (primary objective) is dominated by changes in mass pull, feed PSD and reagent addition. Whereas the particle size to be floated should be rather small and fall within an optimal size range, it has been shown that mass pull is affected by changes in froth stability and structure. At the highest level the flotation primary objective is therefore a function of mass pull, feed

Stellenbosch University http://scholar.sun.ac.za

conditions and reagent addition. Since froth stability and structure are affected by changes in operating parameters, mass pull can be rewritten as a function of air flow rate, froth depth and flotation geometry. From this we can now derive a generic process causality map for flotation (Figure 63).

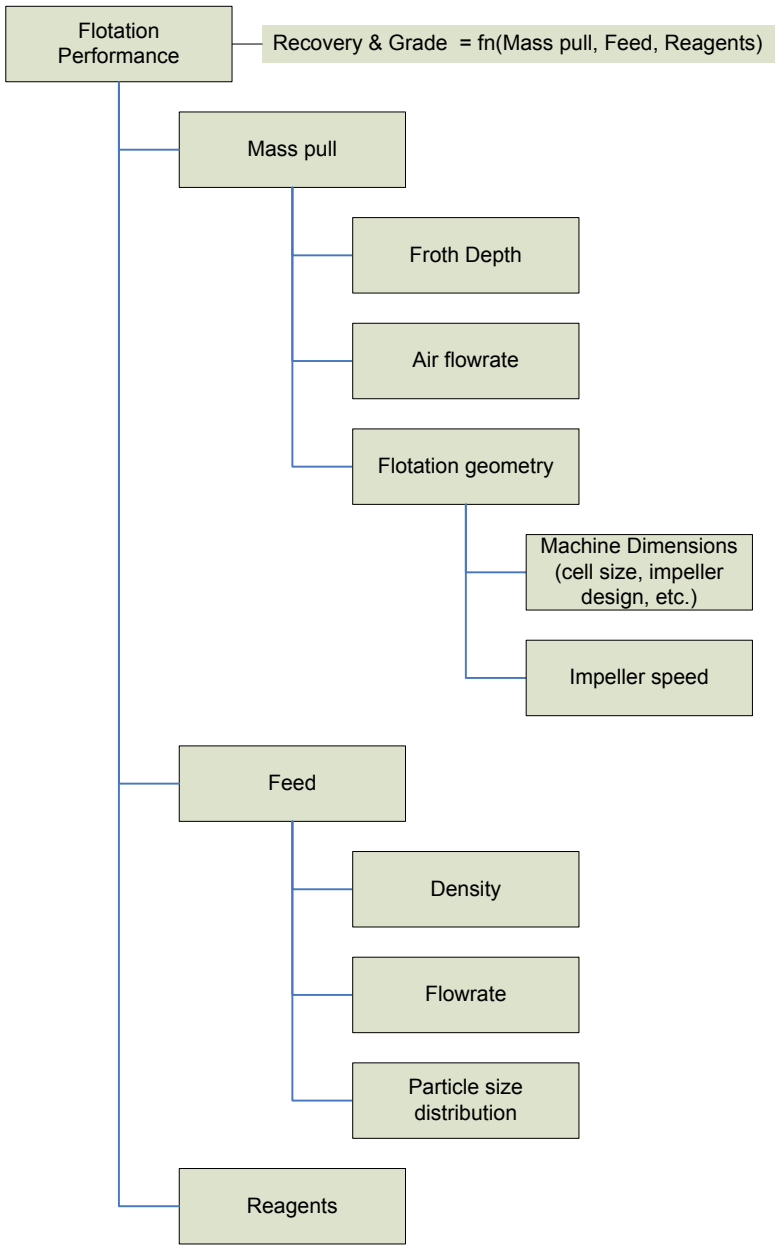

Figure 63: **Generic flotation process causality map**

Stellenbosch University http://scholar.sun.ac.za

6.3.5 RPM(A) No1 UG2 Concentrator

By combining the various process causality maps a view of the entire process can be obtained. For the mineral processing plant under investigation (Figure 57), Figure 64 presents a simplified top level process causality map with regards to process final tails. It is important to note that for each unit operation in this top level process causality map we need to analyse the effect the process had on both the grade and the particle size distribution of the material (where available). This results from the fact that some unit operations primarily affect material grade while others primarily affect material PSD.

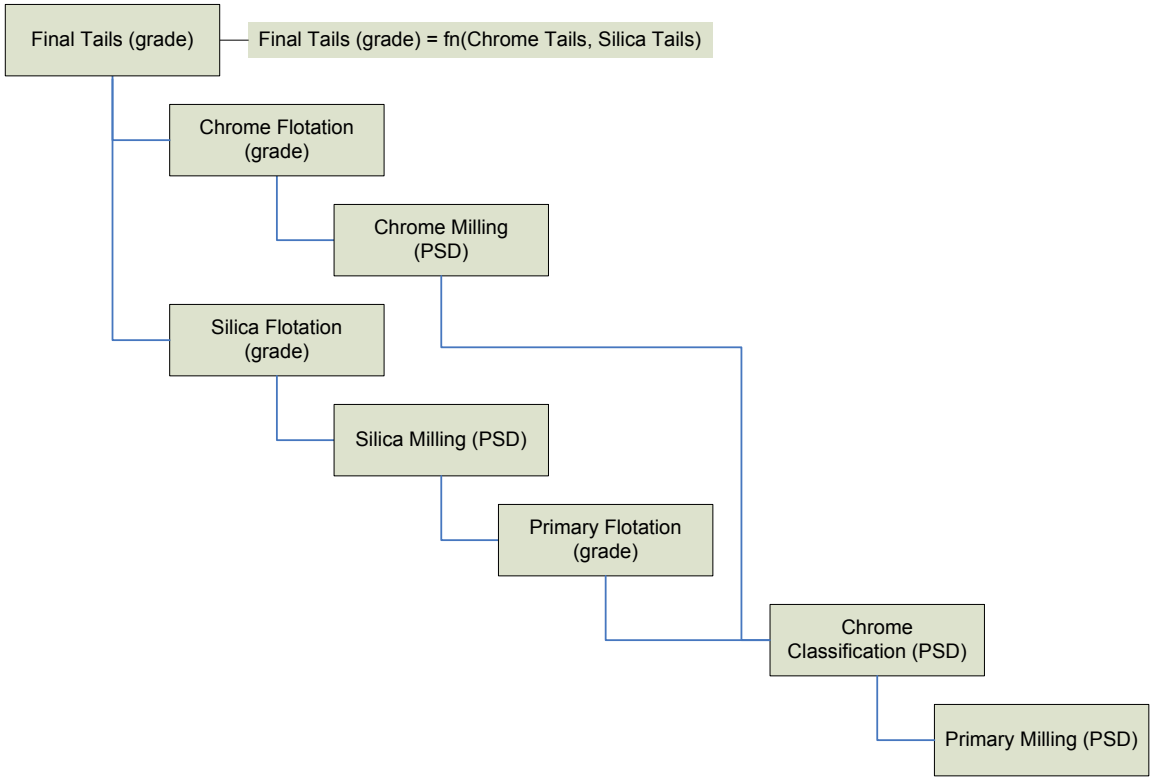

Figure 64: **Generic concentrator process causality map with regards to final tails**

Stellenbosch University http://scholar.sun.ac.za

7 Industrial case study

Having investigated the accuracy, reliability and robustness of the various statistical data-based fault detection, change point detection and variable importance analysis techniques, the application of these techniques, within an analytical framework, to real world problems is required as proof of their usefulness. As part of the case study the importance of not only analysing, optimising and monitoring the process, but also the associated equipment and control system performance will be shown. In all of this, the value of effective data visualisation, making use of graphs such as parallel coordinates, distribution comparisons and 2-D histograms, will be highlighted.

More specifically, this chapter deals with the application of the proposed analytical methodology, with the main focus on the data analysis step, to a mineral processing concentrator case study.

7.1 Concentrator process performance monitoring

The concentrator process performance monitoring case study is based on the concentrator process as described in section 6.2 and shown in Figure 57 (page 180). As is evident from the process description, this is a complex process, with many variables, sampled at different sampling frequencies all of which makes it ideal for analysis using a combination of process causality maps, statistical data-based fault detection, change point detection and variable importance analysis.

Since many events can occur simultaneously within any one time frame in a process, the idea is not to simultaneously focus on all events in isolation but to rather try and understand how different events affect each other and identify the true root cause of a problem. Probably as important as this is the requirement for constructing a view of events showing the interaction between different variables and process operations: not just identifying the root cause of a problem but also showing the cause and effect relationships throughout the process.

It is often easier to encourage people to action by not only showing them what to fix, but also why it is important. This, in part, is why an elaborate analysis is often required for a problem that could potentially be solved in a much simpler way.

7.1.1 Event: Recovery

For this case study, a steady decline was noted in one of the main KPIs of the process: the recovery of the valuable mineral content (a shiftly value that is only available eight-hourly). The decline was visually noted after reviewing a long term trend of the recovery using a four month comparative operational performance monitoring (OPM) report (Figure 65). For the OPM report, the data is arbitrarily divided into two data sets, a reference and current data set. The reference data set spans the first 2 months of the time period under review, with the current data set spanning the last 2 months of the time period under review.

Stellenbosch University http://scholar.sun.ac.za

From the time series trend (Figure 65) it can be seen that as time progresses more of the data starts to exceed the recovery lower warning limit (green dotted line), yellow data markers, and eventually even the recovery lower control limit (red dotted line), red data markers. This is, however, only a visual observation and needs to be substantiated with evidence. As a first pass confirmation of change in the data, a class comparison graph in the form of a box plot is generated (Figure 66). From this it clear that the mean of the data has decreased from between the target value (black line) and the lower warning limit (green dotted line) for the reference data to between the lower warning limit and the lower control limit (red dotted line) for the current data.

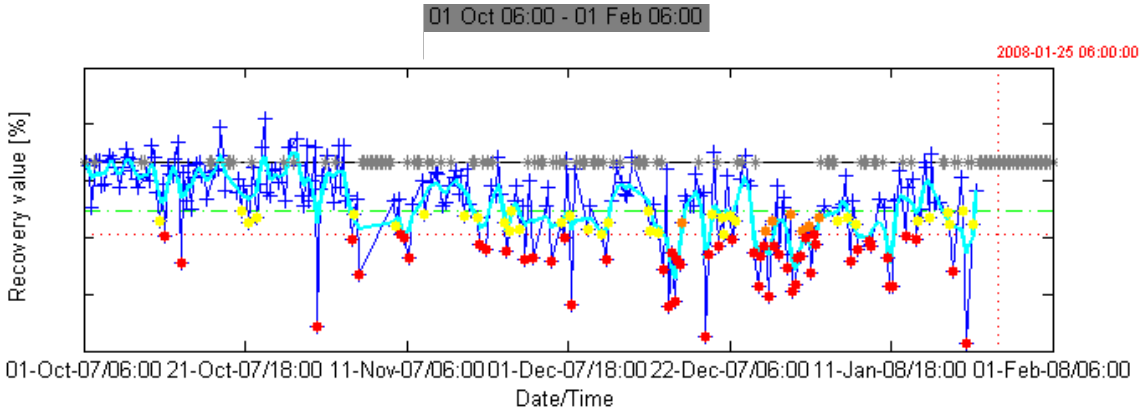

Figure 65: **OPM report: time series trend of recovery**

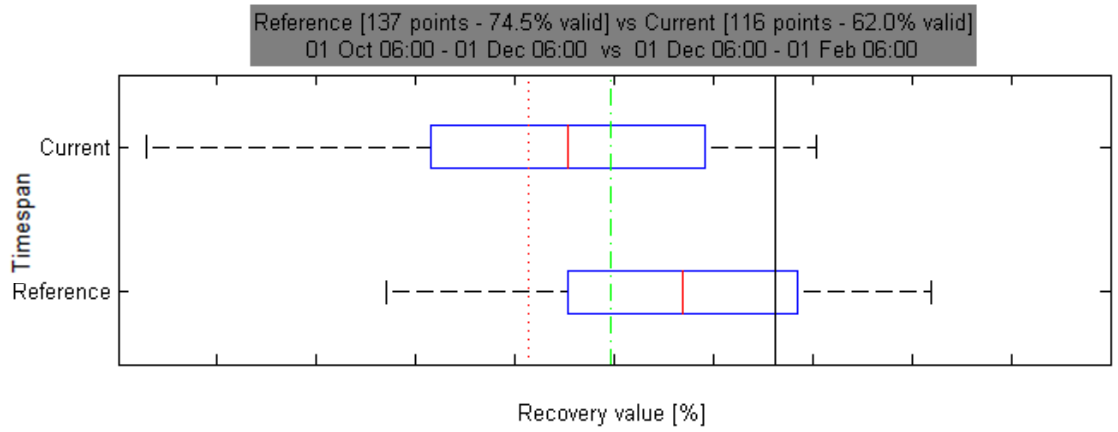

Figure 66: **OPM report: class comparison graph of recovery**

A distribution plot is used (Figure 67) to show the true distribution of the data, displaying both the position and stability of the variable, with normal probability curves fitted to both classes of data simply as a visual aid for visual interpretation of the data. The data is subsequently tested for normality using the Lilliefors test, and in this case was shown to not be normally distributed.

Stellenbosch University http://scholar.sun.ac.za

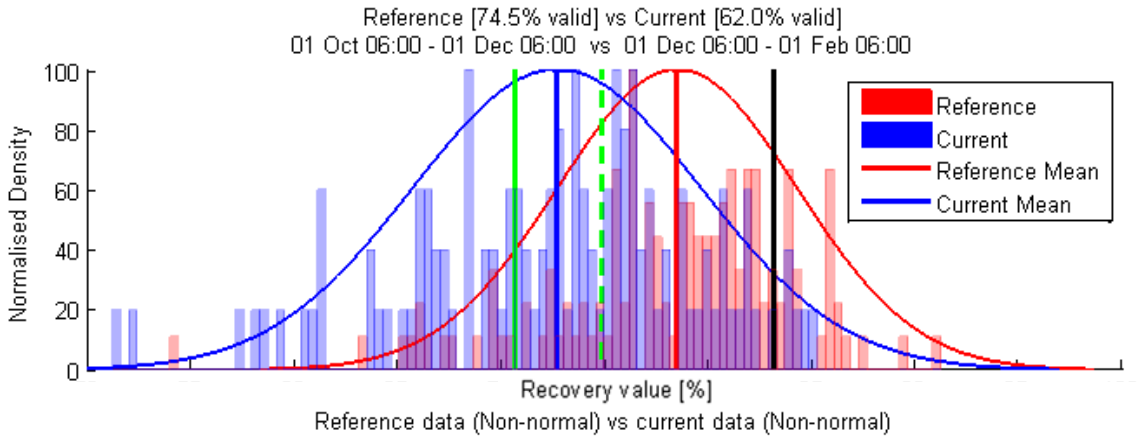

Figure 67: **OPM report: distribution comparison of recovery also indicating the target value (black), lower warning limit (green dotted) and lower control limit (green)**

Prior to a detailed analysis, it is critical to subject all data to data validation and gross error detection. Since data analysis is very reliant on good quality data, data validation is required to determine the quality/trustworthiness of the data. Data validation can be as simple as identifying and removing missing data from a data set to something as complex as a model based data validation approach. The choice of validation techniques and the manner in which the data is handled, removed, used or reconstructed is usually dependant on the type of data analysis that will be performed on the data.

For this analysis, a fairly simple form of gross error detection is performed where the following faulty data is identified:

- Missing data
- Not running data
- High/low data
- Not updating data
- Rate of change exceeded data

For data validation of the recovery variable, the results are presented as a table (Table 11) and graphically through a time series plot of the signal being validated accompanied by a fault graph (Figure 68). In this instance only missing data is detected. This can be ascribed to the origin of the data, where the recovery is a calculated variable dependent on multiple variables automatically sampled, manually processed, analysed and their results captured. Human error is therefore the most likely cause of faulty data in this instance. Data validation is highly dependent on multiple detection thresholds to be entered per variable. These detection thresholds are obtained either from instrumentation specifications or by experience and can be fine-tuned over time.

Stellenbosch University http://scholar.sun.ac.za

Table 11: **OPM report: recovery data validation results**

Ref/Cur	Good [%]	Missing Data [%]	Not Running [%]	High [%]	Low [%]	Not Updating [%]	ROC [%]	General status
Reference	74.46	25.54	0.00	0.00	0.00	0.00	0.00	Bad: Missing Data
Current	62.03	37.97	0.00	0.00	0.00	0.00	0.00	Bad: Missing Data

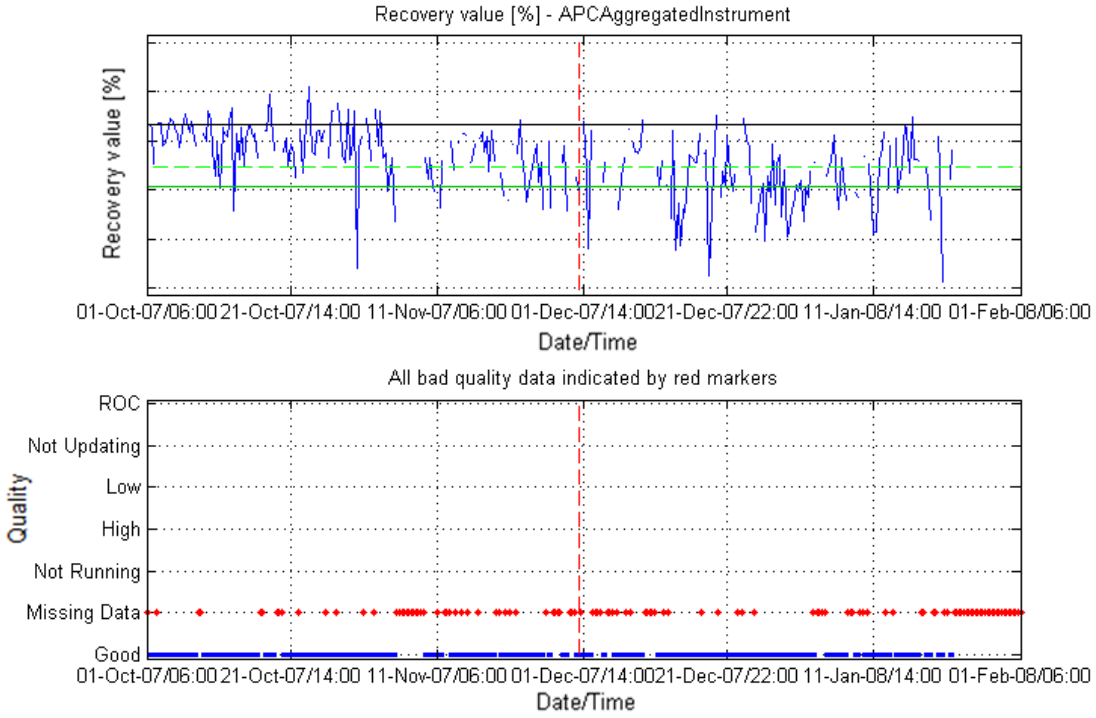

Figure 68: **OPM report: recovery data validation graph**

As identified, the recovery variable data set contains a fair amount of missing data (Table 11) for both the reference and current data classes requiring missing data replacement to be looked at prior to data analysis. For this analysis, missing data was replaced with the mean (together with some additive Gaussian noise with a low variance) of the non-missing data over the entire data range. If this is found not to be effective, alternative, more advanced missing data replacement approaches could be applied. One such a technique makes use of auto-associative neural networks.

An auto-associative neural network contains information that allows replacement of missing sensors with values estimated from remaining sensors. Two assumptions are needed: (1) input variables are correlated; (2) the training data set contains enough information to cover normal process operation. One approach to follow is that the most likely value for the missing sensor is the value that minimizes the magnitude of the deviation between the input and output vectors. For a single fault sensor, the problem is an univariate optimization problem and during the optimization, the values of the remaining sensors are fixed at their measured values. For more than one faulty sensor at a time the problem becomes a multivariable problem.

Following the replacement of the missing data, shown as red on the recovery variable time series trend in Figure 69, various data characteristics are estimated in order to determine the validity of the change point detection techniques to be used:

- Data is auto-correlated at a 99.97% significance level
- Data is not normally distributed at a 95% significance level
- Data is not exponentially distributed at a 95% significance level

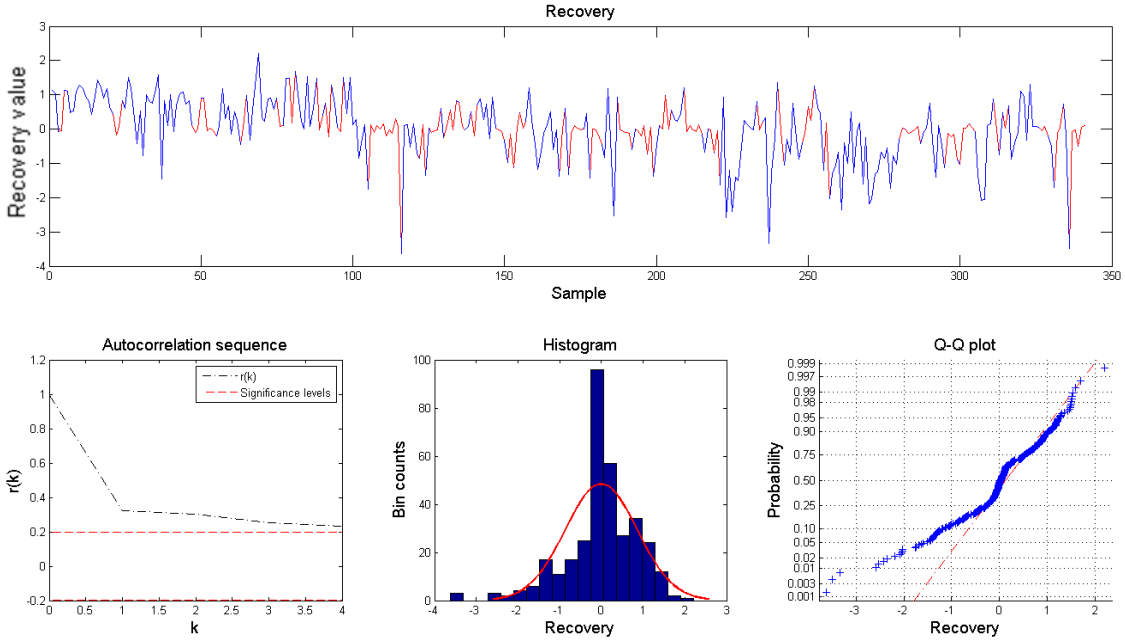

Figure 69: **Data characterisation: recovery**

Following data characterisation, potential change points in the data needs to be identified. Although the data characteristics shows that the SSA change point detection technique is the most suitable, it was shown during the technique evaluation that the Bayesian change point detection algorithm was very robust irrespective of the data characteristics and proved to be the most reliable, and thus the preferred

Stellenbosch University http://scholar.sun.ac.za

algorithm whose results will be considered more favourable than the other change point detection algorithms.

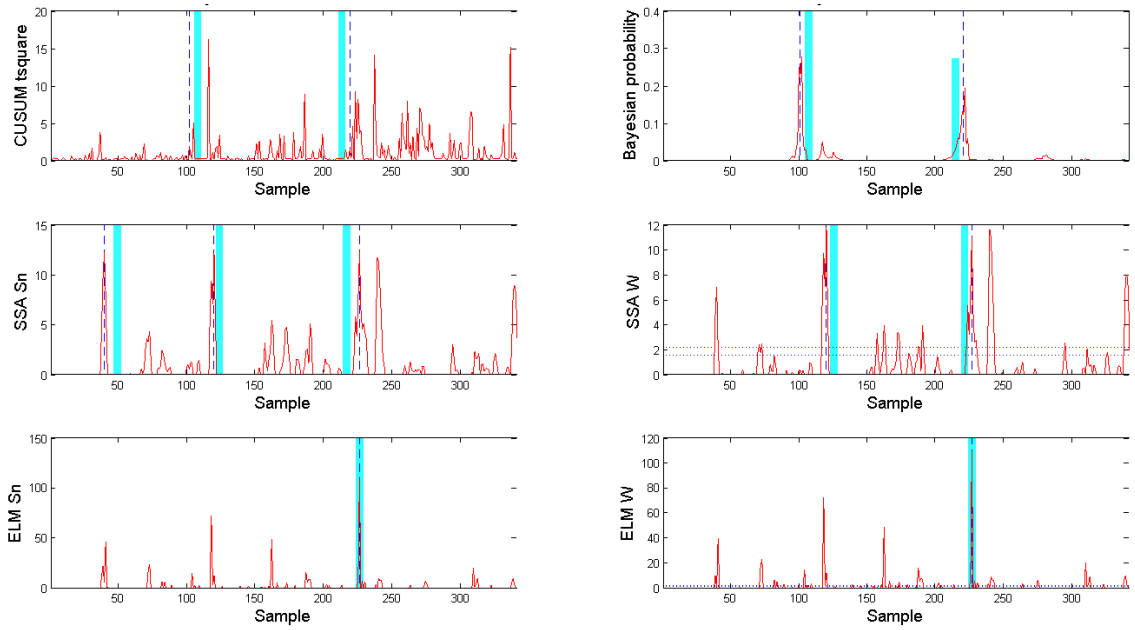

Figure 70: **Change point detection: recovery at a confidence level of 0.99**

From the change point detection results (Figure 70), especially considering the Bayesian probability, it is evident that change points occur at sample 101 (08/11/2007) and 220 (13/12/2007). Both these change points also coincide with what appears to be relevant change points detected by the nearest-neighbours CUSUM algorithm. In contrast to this, the both the SSA and ELM SSA algorithms identified other seemingly important change points. However, based on the technique evaluation findings, the focus will be on the change points identified by the Bayesian probability technique.

Using these identified change points, the data is classed into three distinct classes:

1. Class 1 – reference data – data points 1-100
2. Class 2 – changeover data – data points 101-219
3. Class 3 – fault data – data points 220-341

Subjecting these 3 recovery classes to the median significance (Figure 71) and one-way ANOVA analysis, it was found that there was a significant difference between the classes at a significance level of less than 0.05. The decrease in recovery was subsequently confirmed by plant personnel; unfortunately no reason as to the cause thereof could be given.

Stellenbosch University http://scholar.sun.ac.za

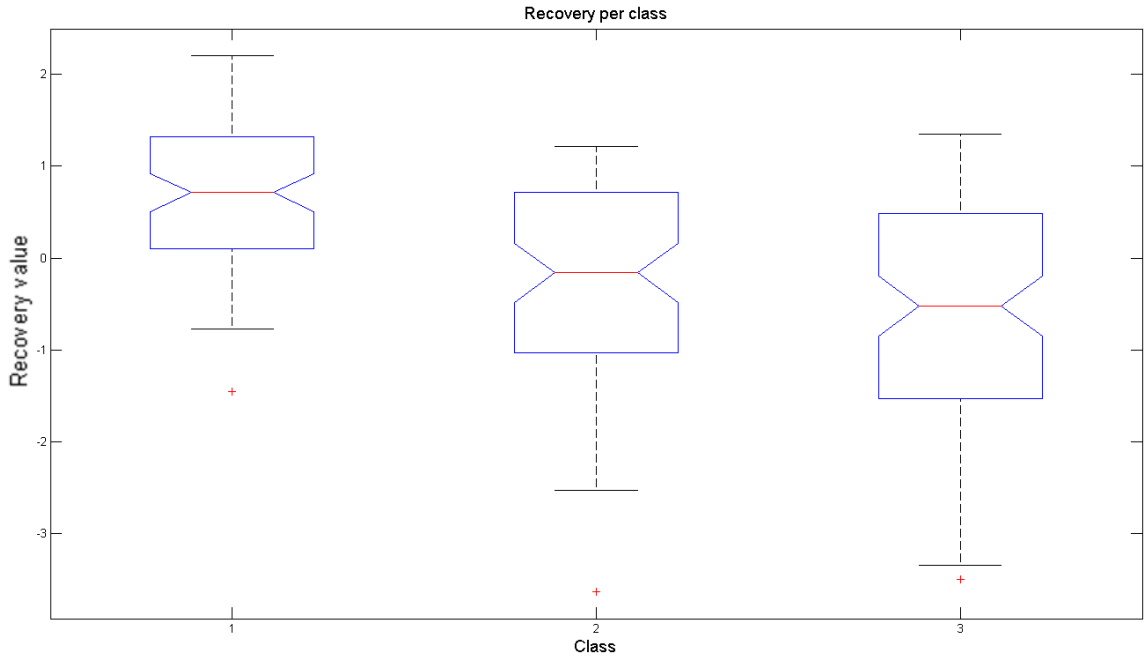

Figure 71: **Median significance: recovery**

Now that it has been confirmed that a change did in fact occur in the process, the analysis needs to be initiated with proper problem identification and data collection, followed by a thorough understanding of how the decrease in the recovery was brought about.

7.1.2 Analysis preparation

With the event detected in one of the main KPIs in the output of the concentrator process, the boundary for the analysis needs to effectively be drawn around the entire concentrator process as per the process flow diagram in Figure 57 and the process causality map in Figure 64. The analysis will focus on the three data classes as determined by the change point detection performed on the recovery variable. Although more change points may be identified in the data as the analysis progresses and different variables are analysed, the data will not be re-classed although this could be done to further improve the findings of the case study. Variables and KPIs will be analysed as per those defined for the various process causality maps.

Following the problem identification, relevant data need to be collected. Typical relevant data that would be required either directly for analyses purposes or for results interpretation include:

- Process flow diagram
- Piping and instrumentation diagrams – these not only show what instrumentation is available and where it is located in the process, but generally can also be used to identify the tag names for retrieving the process data from the process historian

Stellenbosch University http://scholar.sun.ac.za

- Laboratory analysis results for feed, intermediary, concentrate and tailings streams – this data is typically available shiftly and/or daily. Analysis includes:
  - PGM content
  - Base metal content
  - Grind
- Online process data – this data is typically available at a sampling frequency of one every ten seconds or less. Measurements include:
  - Process operating measurements such as flows, densities, levels, etc.
  - Equipment operating measurements such as run, amps, power, etc.
  - Controller measurements such as set points, outputs, auto/manual, etc.
  - Online stream analyser measurements.
- Derived measurements and KPIs – this data is typically available at the lowest frequency of the composite measurements used to derived the values. Measurements include:
  - Mill stops
  - Equipment availability
  - Mass split
- General measurement quality information – this data is typically diarised:
  - If laboratory measurements, were analysis techniques changed?
  - If online stream analyser measurements, were recalibrations done?
  - If process measurements, were instrument recalibrations done?
- General equipment configuration change information – this data is typically diarised:
  - If grizzlies/screens, were classification sizes changed?
  - If mills, were mills relined?
  - If ball mills, obtain ball loading schedule?
  - If cyclones, were physical changes made to cyclones?
  - If cyclone clusters, were number of cyclones in operation changed?
  - If float, were float configuration changed?
  - If spillage, was spillage configuration changed?

7.1.3 Analysis overview

As a first pass analysis, all the variables, 56 in total, are analysed simultaneously. Because of the different sampling rates between different groups of variables, the data is re-sampled for all the variables to create a data set consisting of shiftly (8 hourly) data. The data set is first subjected to change point detection analysis, followed by fault detection and variable importance analysis.

For the combined data set (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), none of the change point detection algorithms seemed able to reliably detect any significant change points in the data set (Figure 72). Visual inspection of the nearest-neighbours CUSUM algorithm do, however, indicate that changes occurred in the data set in the vicinity of samples 100 and 200, with the latter being more pronounced.

Stellenbosch University http://scholar.sun.ac.za

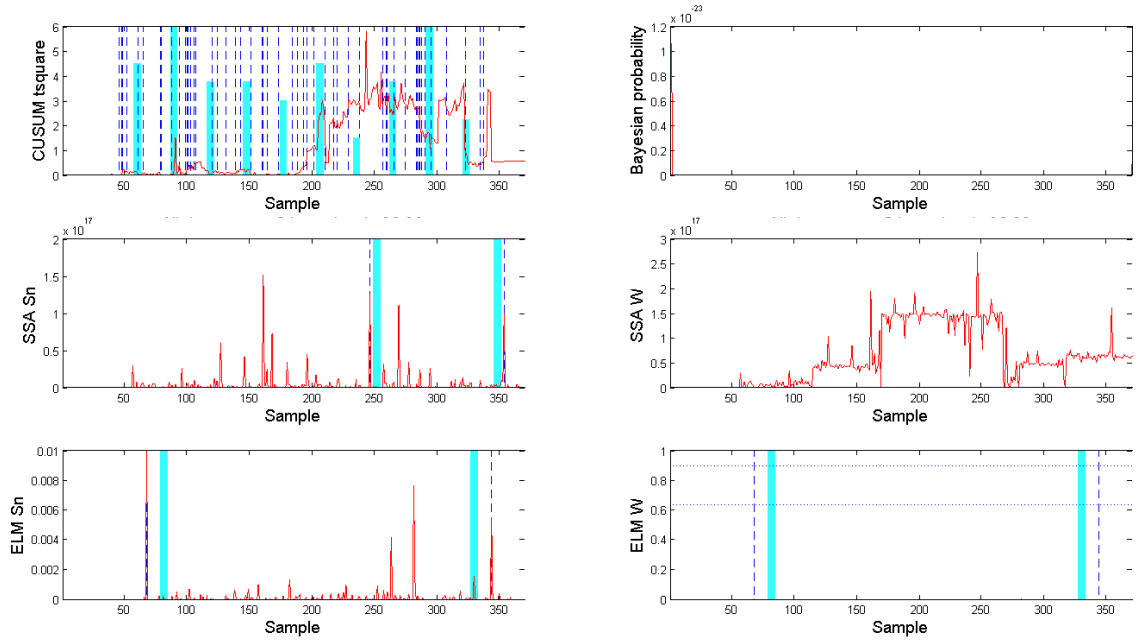

Figure 72: **Change point detection: all data at a confidence level of 0.99**

Using only the class 1 reference data from the combined data set, fault detection models were constructed to be used for the future detection of similar fault conditions. For these models it was found that the CUSUM performance metric for some of the variables had very high false alarm rates (Figure 73), potentially rendering this performance metric unsuitable for fault detection. Subsequently, the performance metrics were calculated for the class 2 changeover and class 3 fault data. For the reliability index (Figure 73) it was found that the non-linear multivariate performance metrics performed the best, followed by the dynamic multivariate performance metrics, the basic multivariate performance metrics and lastly the univariate performance metrics. This highlighted the fact that the data set is highly non-linear and dynamic. Of the performance metrics having high reliability, the CUSUM and summed-scores performance metrics also had relatively large detection-delays (Figure 73) which, for the CUSUM performance metric, confirmed its unsuitability for fault detection. It can therefore be concluded that the majority of the multivariate performance metrics were able to (and should be able to for similar future fault conditions) reliably distinguish between the reference data and the changeover and fault data, also confirming the classes identified by the change point detection performed on the recovery variable.

Stellenbosch University http://scholar.sun.ac.za

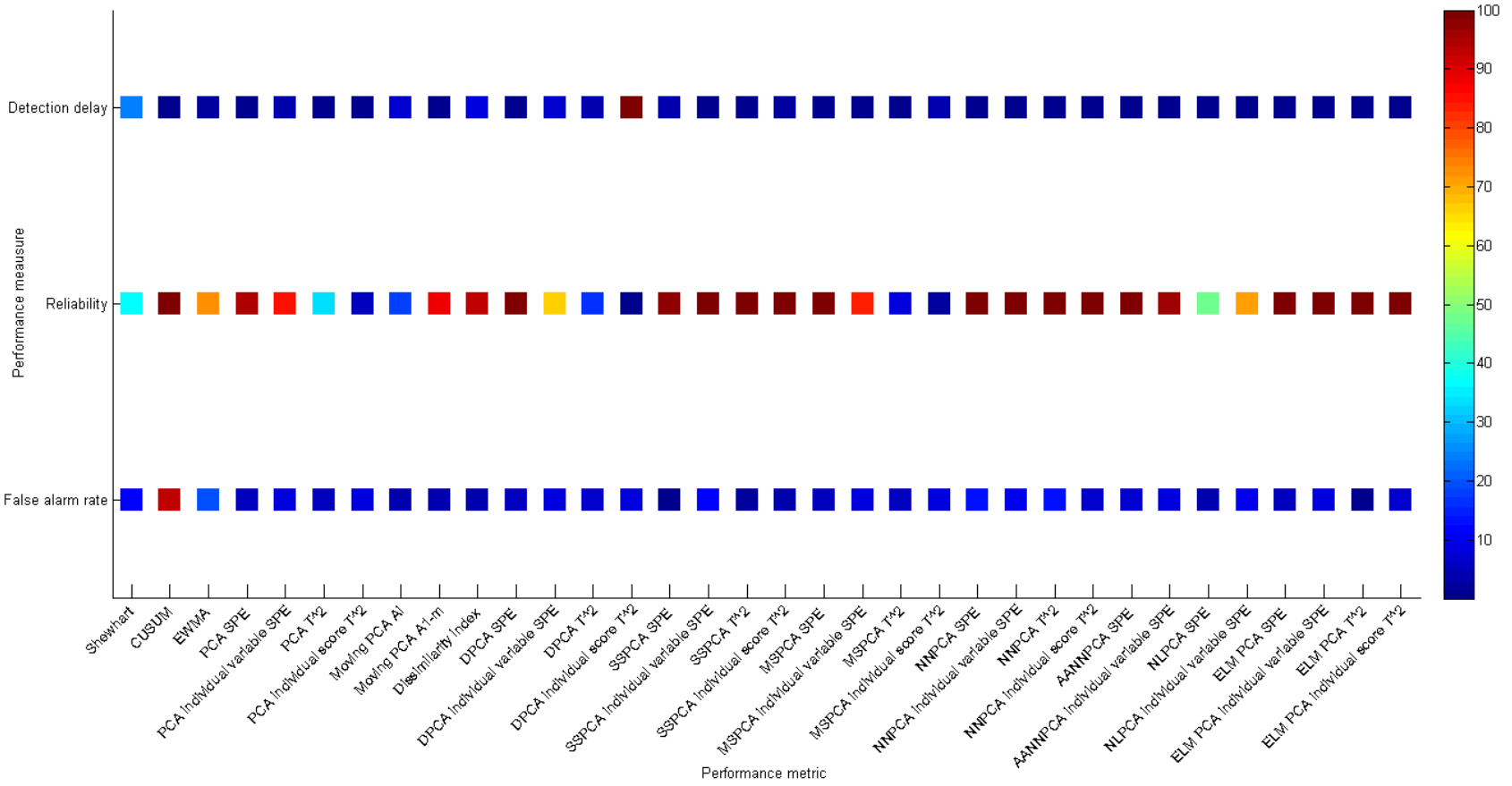

Figure 73: **Statistical data-based fault detection: all data – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 74) is made to determine if the combined variables can be used to distinguish between the different recovery classes.

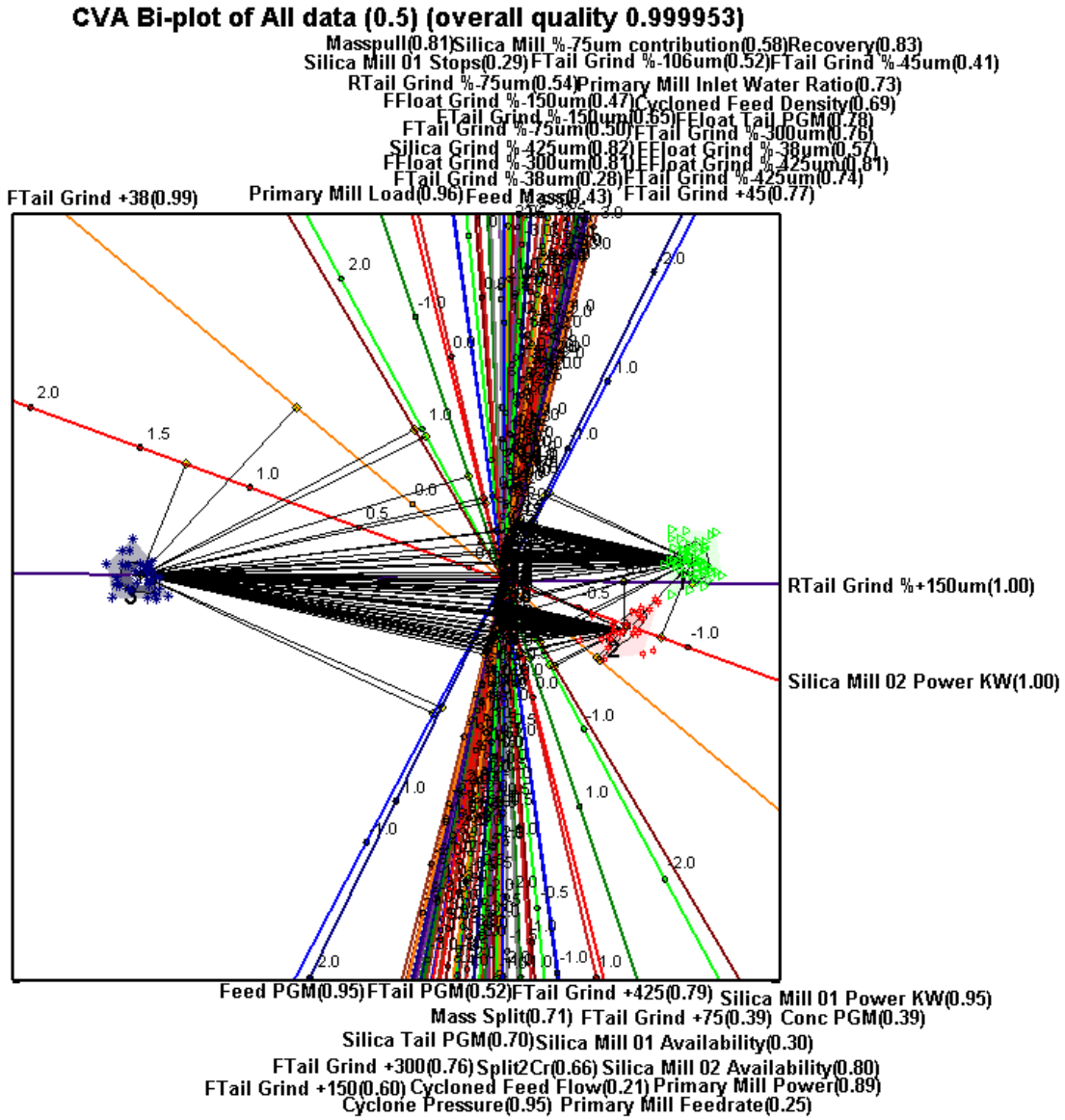

Figure 74: **CVA biplot illustrating the overly cluttered nature of the graph when analysing too many variables simultaneously: all data (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

It can be seen from the CVA biplot that although the classes are very well separated, with the alpha bags for each class not overlapping each other at all, the graph is overly cluttered with regards to variable axes. Furthermore, the overall quality of the data projection is very high, with the variable axes having

Stellenbosch University http://scholar.sun.ac.za

predictivities ranging from very low (removed from the graph) to very high. Due to the cluttered nature of the CVA biplot, very little can be said with regards to the correlation between the variables analysed.

Visually, the only identifiable variables that can easily be associated with their respective axes, are the silica mill 02 power, the primary flotation grind (RTail grind) and the final tail grind (FTail grind). This is, however, not very informative as many of the other variables are also bound to represent important information, but this cannot be sensibly gained from the CVA biplot when representing all 56 variables.

The CVA biplot analysis is supported by the variable importance analysis consisting of the random forests, classification trees and ELM-with-bagging techniques (the linear discriminant variable importance technique was not applied due to there being too few data per class relative to the number of variables). For ease of interpretation, the individual prediction results from these techniques were normalised and summed to produce a single variable ranking for the recovery drivers (Figure 75, Figure 76 and Figure 77). Individual prediction results for the specific techniques are, however, still displayed to assist with the results interpretation.

From the variable importance analysis on average over all 3 classes the main differentiators between the classes are the silica mill 02 power, the chrome classification cyclone feed flow, the final tail grind (PSD) and the silica mill 01 power. Moving only from class 1 to class 2, the chrome classification cyclone feed flow rate variable became the most important differentiator, followed by the final tail grind, silica mill 1 power and the chrome classification mass split. Moving only from class 2 to class 3, these differentiators reverted back to 3 of the top 4 identified for all 3 classes, including chrome classification cyclone pressure and the primary mill power.

Stellenbosch University http://scholar.sun.ac.za

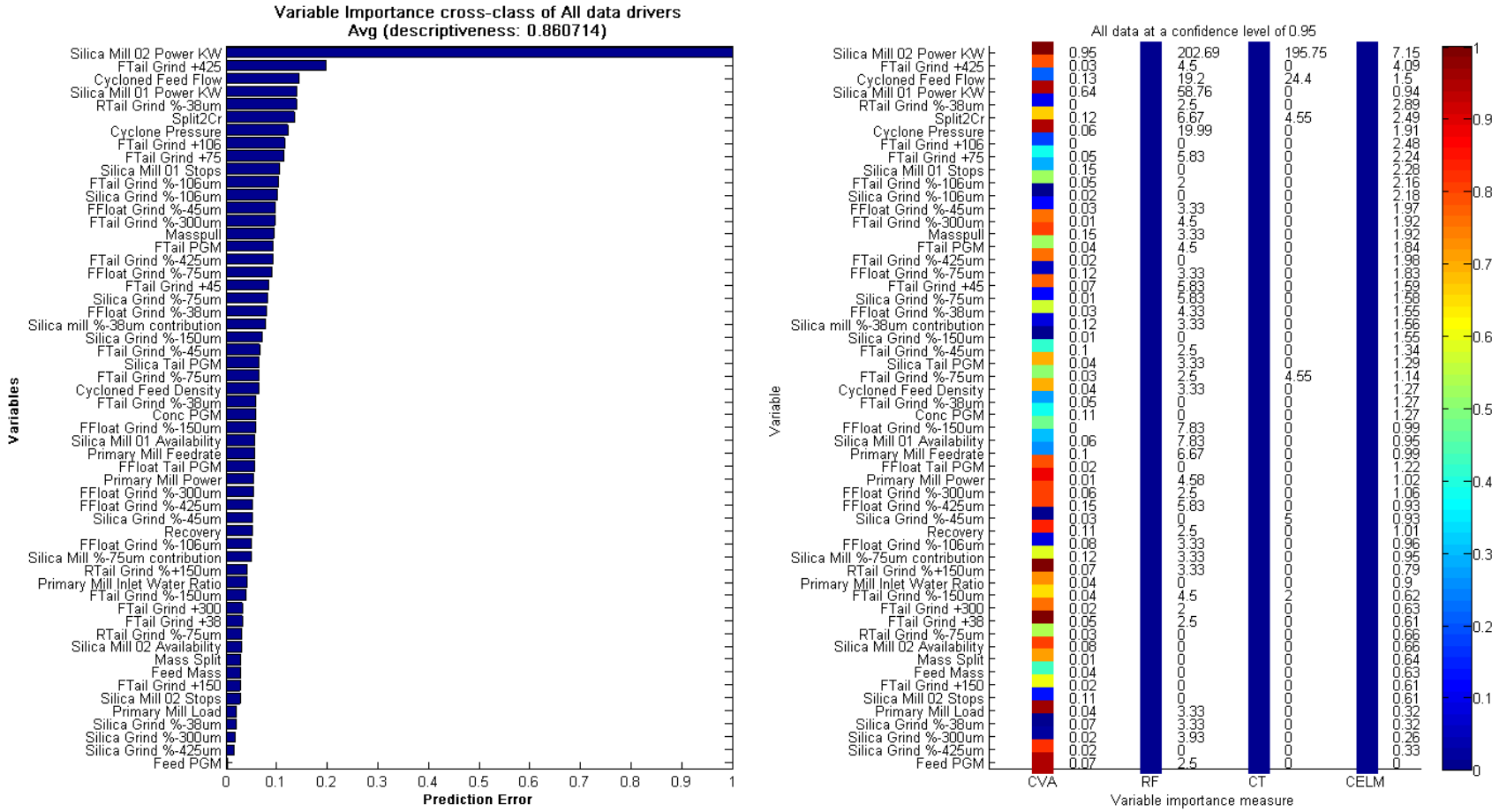

Figure 75: **Variable importance class 1-2-3: all data at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

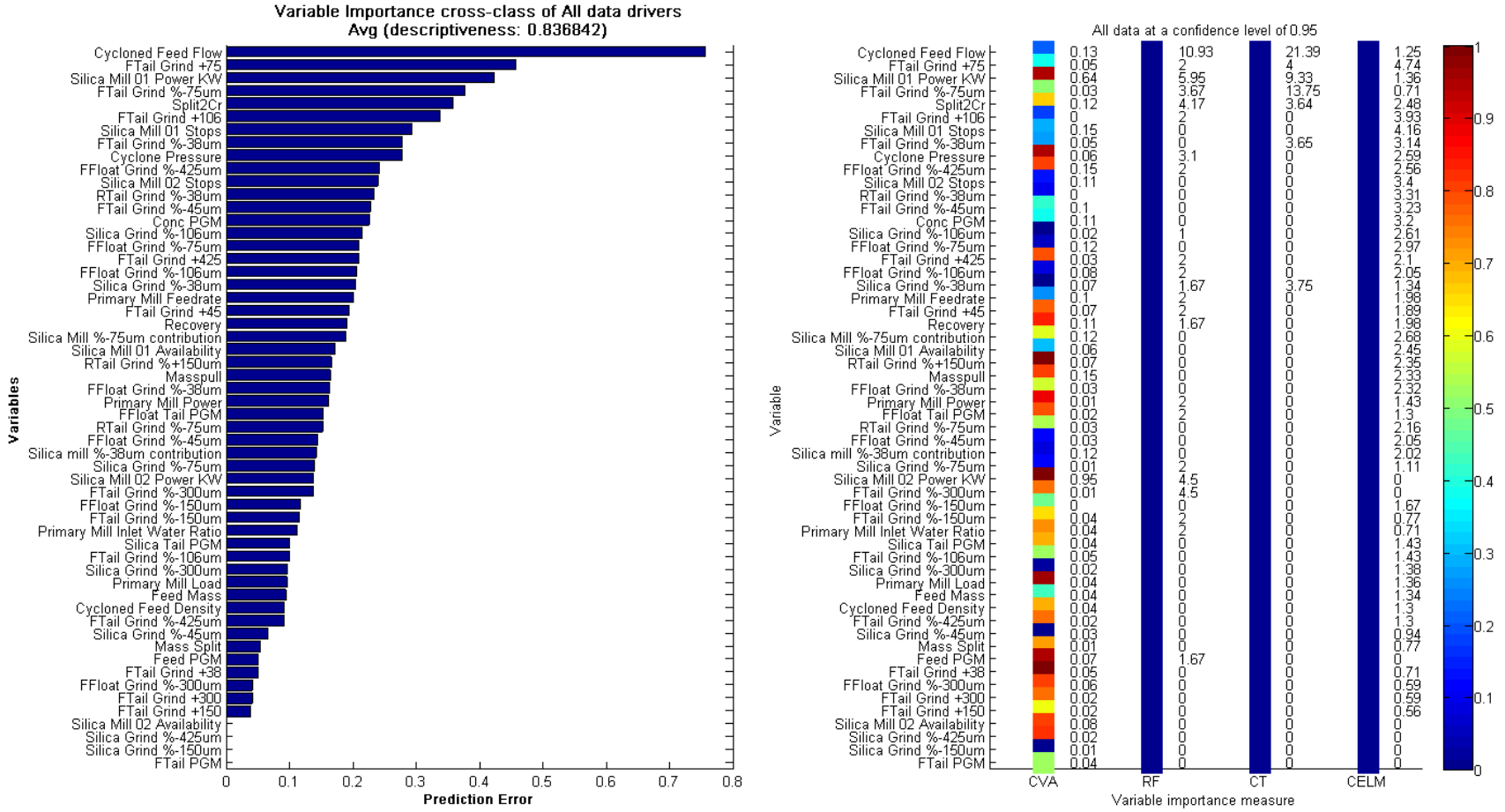

Figure 76: **Variable importance class 1-2: all data at a confidence level of 0.95**

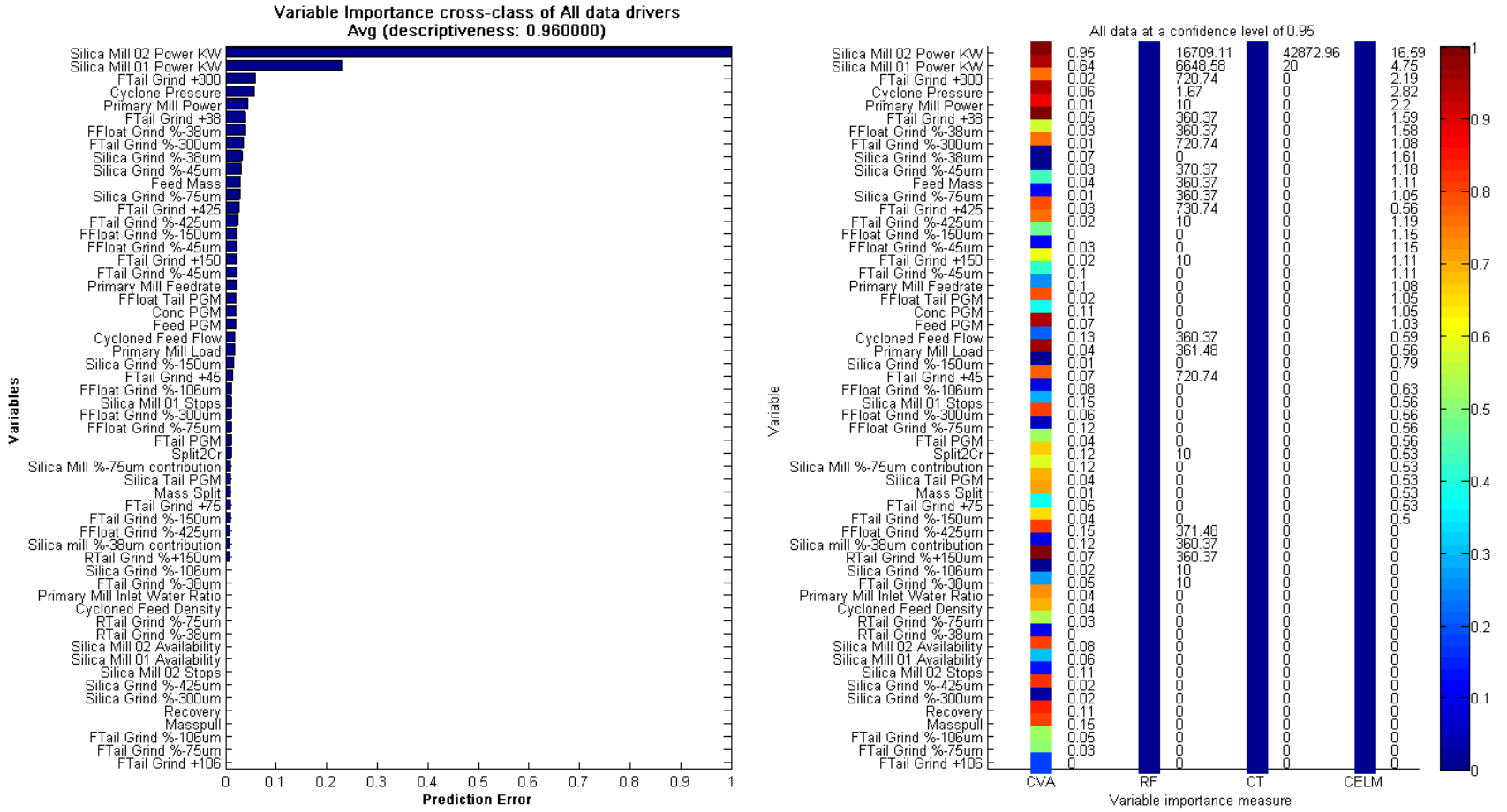

Figure 77: **Variable importance class 2-3: all data at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Although many events have been highlighted and important variables identified, nothing can be said about the causality between these events and the variables or regarding the root cause of the decrease in recovery. As was highlighted during the technique evaluation for the Tennessee Eastman process, for complex processes having many variables requiring monitoring, the complexity of the problem could be greatly reduced through the application of process causality maps. Process causality maps will allow multiple, smaller, individual processes to be monitored at a low level, while still allowing the overall process to be monitored at a higher level.

7.1.4 Drivers: Recovery

With the analysis prepared, and an overview analysis performed, the focus shifts to incorporating process causality maps in the search for finding the drivers that have caused the decrease in the recovery. From a process causality point of view, the recovery is a function of plant feed, final concentrate produced and final tails discarded. The process streams are represented (where available) by the following measurements:

1. Plant feed (Feed)
    a. Plant feed head grade
    b. Plant feed tons treated
    c. Feed PSD
2. Final concentrate (Conc)
    a. Final concentrate grade
    b. Mass pull
3. Final tails (FTail)
    a. Final tails grade
    b. Final tails PSD

As before, fault detection models were constructed using only the class 1 reference data from the recovery drivers data set with performance metrics calculated for the class 2 changeover and class 3 fault data. Considering this data set is only a subset of the combined data set, the univariate performance metrics delivered the exact same results (Figure 78), and therefore conclusions, as for the evaluation of the combined data set. For the reliability index (Figure 78) it was again found that the non-linear multivariate performance metrics performed the best, followed by the dynamic multivariate performance metrics, the basic multivariate performance metrics and lastly the univariate performance metrics. It should, however, be noted that for the recovery drivers data set, less performance metrics were reliable when compared to the combined data set. This alludes to the fact that some of the other process variables are probably better indicators of a fault condition in the process compared to the process variables in the recovery drivers data set. Again a large number of the multivariate performance metrics were able to reliably distinguish between the reference data and the changeover and faulty data, confirming the classes identified by the change point detection performed on the recovery variable.

Stellenbosch University http://scholar.sun.ac.za

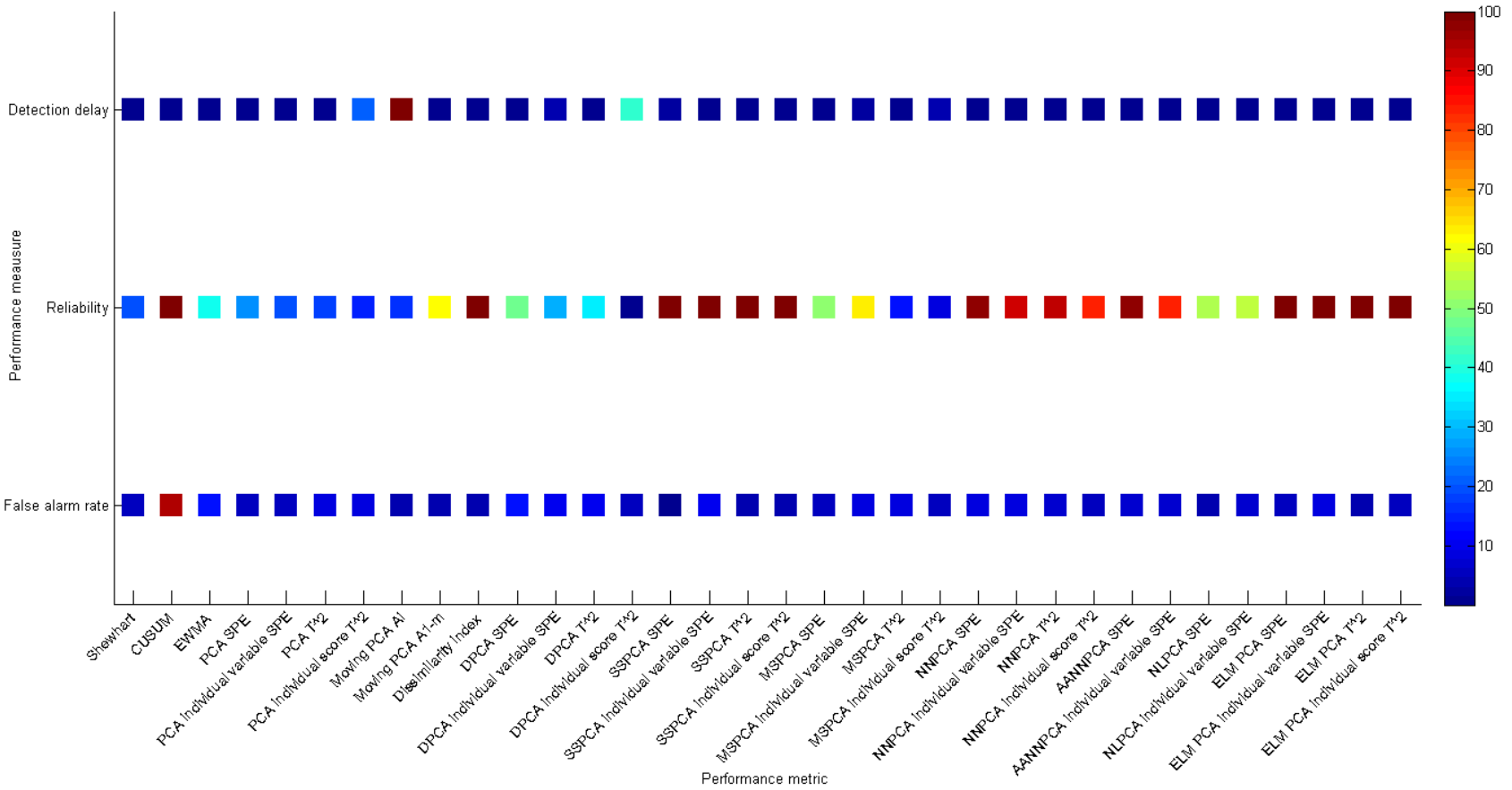

Figure 78: **Statistical data-based fault detection: recovery – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 79) is made to determine if the recovery drivers can be used to distinguish between the different recovery classes.

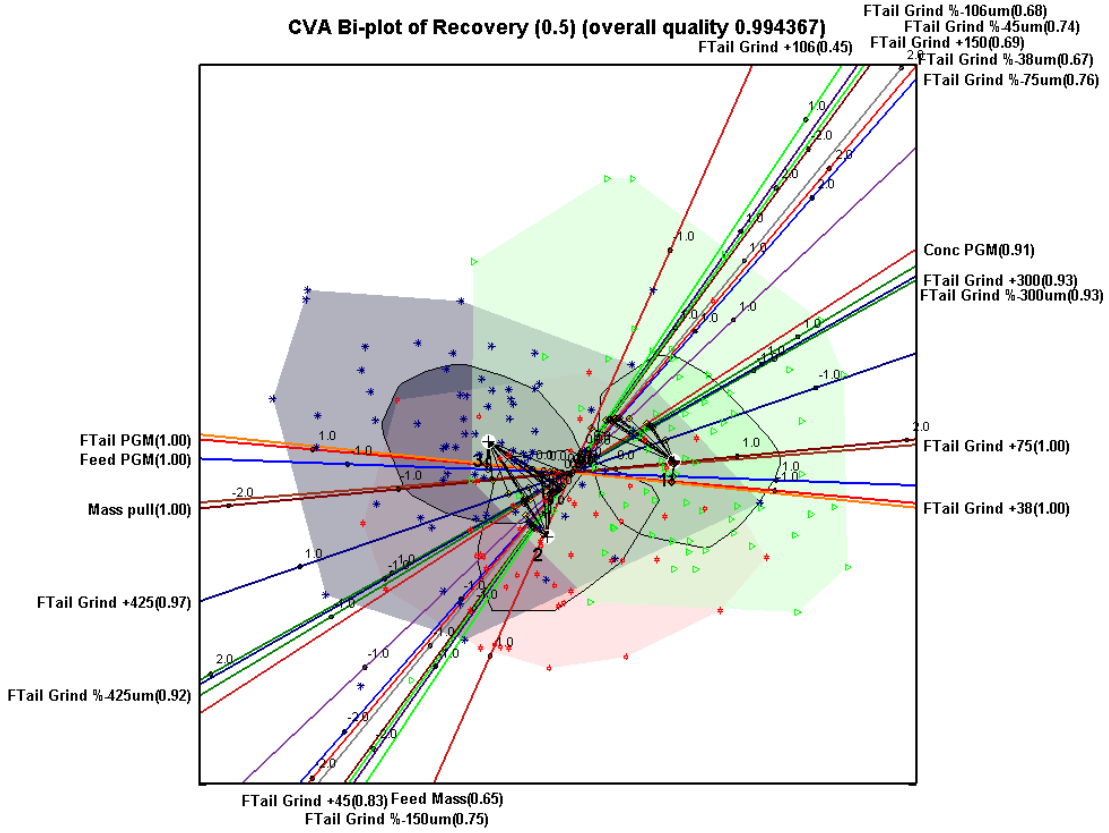

Figure 79: **CVA biplot: recovery (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

It can be seen from the CVA biplot that the classes are not well separated as is evident from the fact that there is overlap of the alpha bags, even at an alpha value of 0.5. The overall quality of the data projection is, however, very high at 0.99 (1 being the maximum), with most variable axes also having high predictivities, close to or equal to 1, in the 2-dimensional projection. The conclusions on the discriminatory ability of most of the variables from this projection should therefore be significant and reliable.

As expected, it can be seen from the CVA biplot that there is a high degree of correlation between most of the final tails grind measurements, as is evident from the small angle between the biplot axes for these measurements. A similar, and due to the nature of the process also expected, high degree of correlation is found between the plant feed head grade, the final tails grade and the mass pull measurements. Visually, it seems that on average over all 3 classes the main differentiators between the classes are final tail grade (together with the final tail grind +75 μm and +38 μm measurements), plant feed head grade

and mass pull (all also having axes predictivities of 1). However, when moving only from class 1 to class 2, these drivers are overshadowed by the final tail grind (PSD) and the plant feed tons treated.

The CVA biplot analysis is supported by the variable importance analysis consisting of the random forests, linear discriminant, classification trees and ELM-with-bagging variable importance techniques. From the variable importance analysis on average over all 3 classes (Figure 80) the main differentiators between the classes are mass pull, final tail grind (PSD), plant feed head grade and final tail grade. As with the CVA biplot analysis, different variables stood out as important when considering only 2 classes at a time. Moving only from class 1 to class 2 (Figure 81), final tail grind (PSD) and mass pull was the most important drivers with the plant feed head grade being the most important driver for the move from class 2 to class 3 (Figure 82).

Although the mass pull measurement seems to stand out as the single most important variable when considering a shift between all 3 classes, it may be masking more fundamental shifts in the process that occur when moving between class 1 and class 2 and between class 2 and class 3 in the data. Further analysis will, therefore, not only focus on the important drivers over all classes (mass pull), but also on the drivers between individual classes (final tails grind for movement between class 1 and class 2 and plant feed head grade for movement between class 2 and class 3).

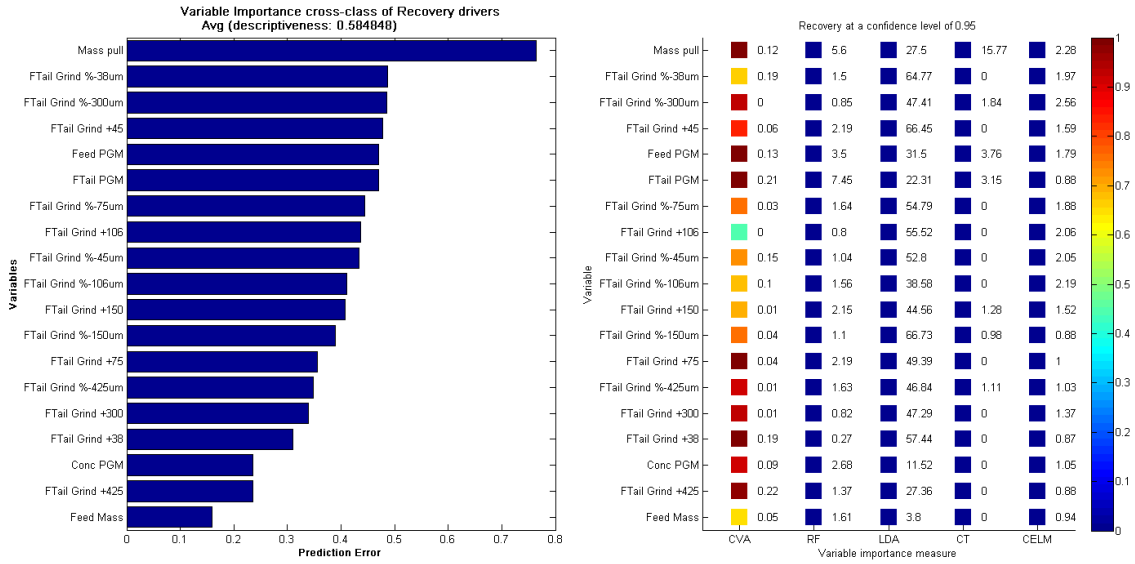

Figure 80: **Variable importance class 1-2-3: recovery at a confidence level of 0.95**

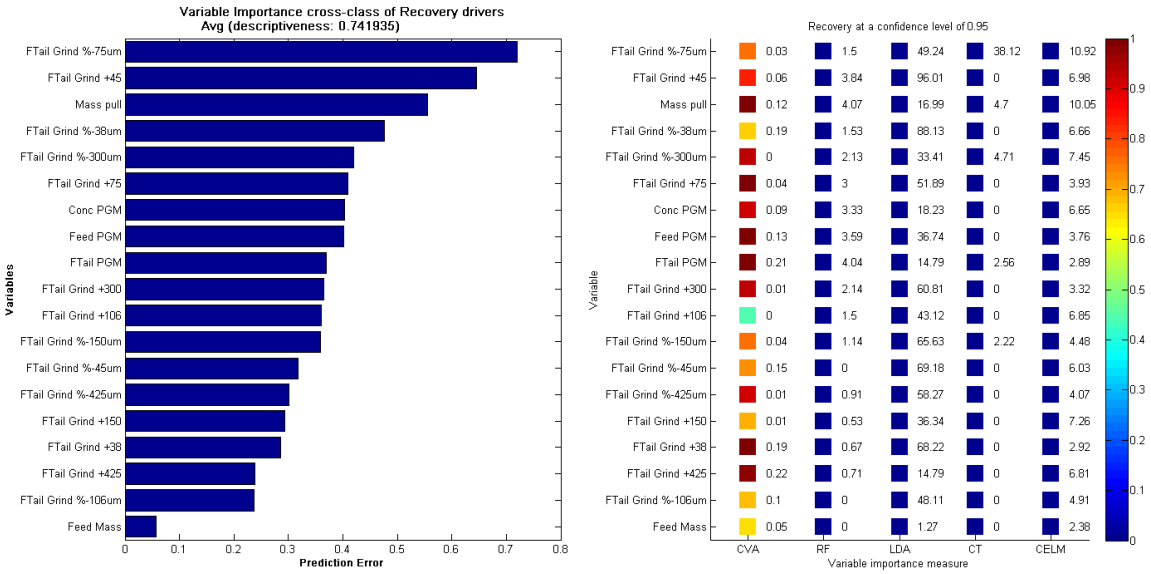

Figure 81: **Variable importance class 1-2: recovery at a confidence level of 0.95**

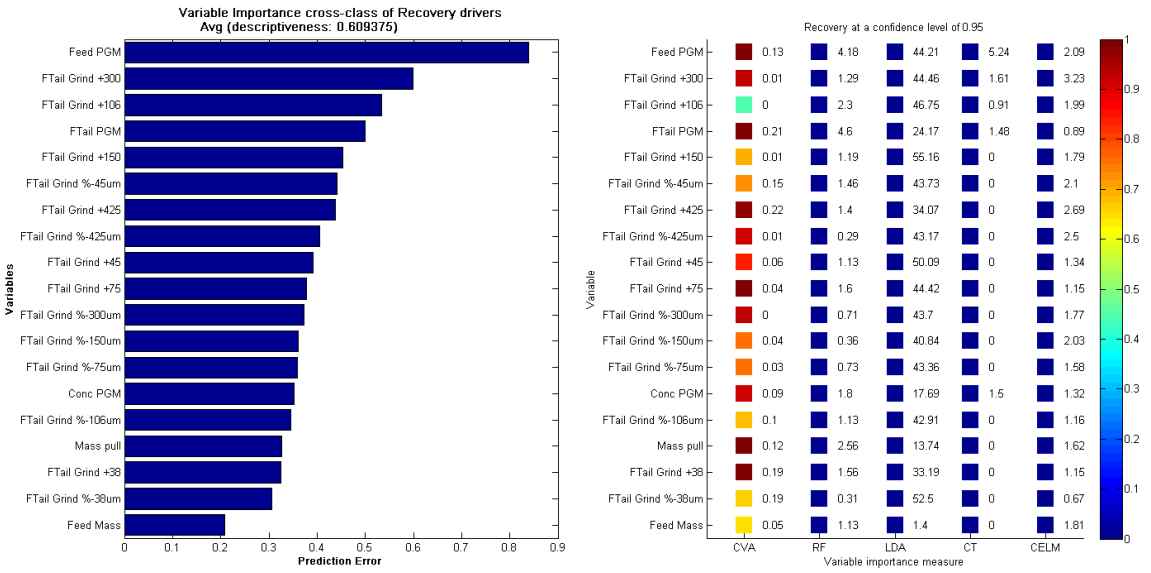

Figure 82: **Variable importance class 2-3: recovery at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

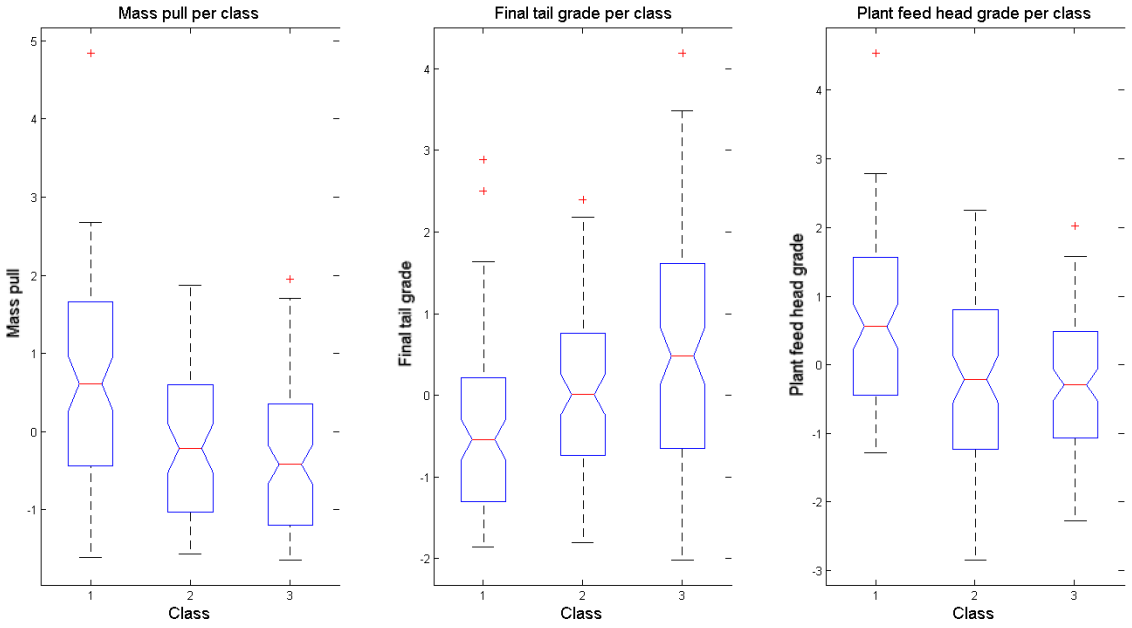

Figure 83: **Median significance: mass pull, final tail grade and plant feed head grade**

Subjecting the 3 mass pull, final tail grade and plant feed head grade classes of data to the median significance (Figure 83) and one-way ANOVA analysis, it was found that there was a significant difference between the classes for all measurements at a significance level of less than 0.05. Similarly, for the final tails grind, for all size fractions there is a significant difference between the classes at a significance level of less than 0.05 (Table 12), with the final tail grind becoming noticeably coarser before improving once again (Figure 84).

Table 12: **Grind significance: final tails grind**

Variable	Significance value
Final Tail Grind %-38 µm	0.0112
Final Tail Grind %-45 µm	0.0051
Final Tail Grind %-75 µm	1.3756e-005
Final Tail Grind %-106 µm	2.1297e-006
Final Tail Grind %-150 µm	5.6610e-006
Final Tail Grind %-300 µm	8.9390e-006
Final Tail Grind %-425 µm	1.5450e-004

Stellenbosch University http://scholar.sun.ac.za

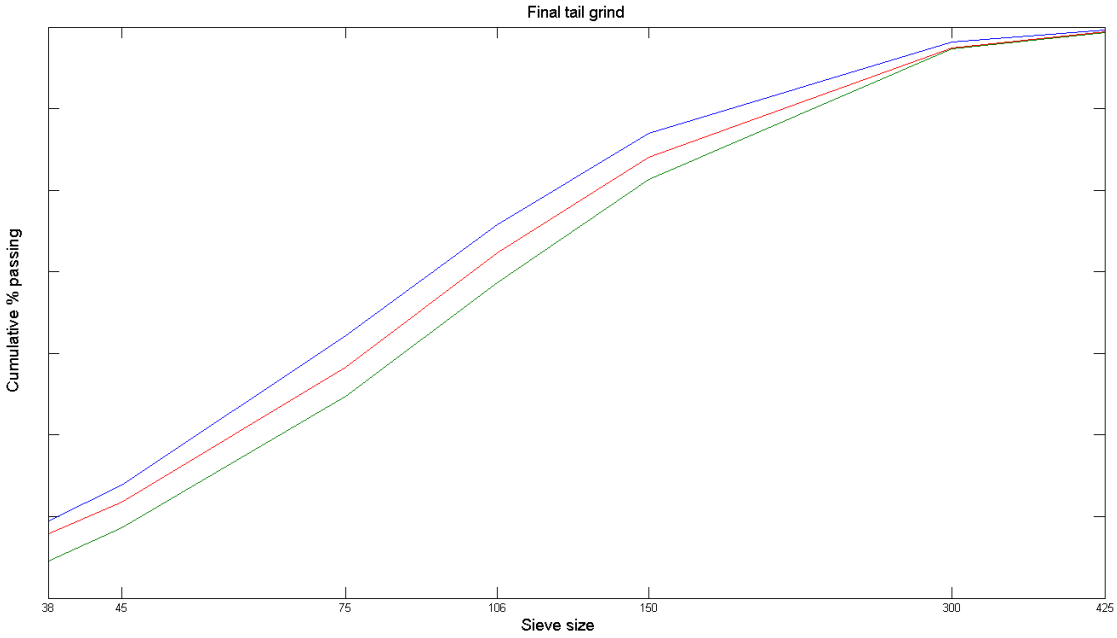

Figure 84: **Cumulative % passing grind curve: final tails grind (class 1 – blue; class 2 – green; class 3 – red)**

Although a good correlation has been found thus far between the change points detected in the recovery variable and potential important variables, it would be interesting to see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others.

For the mass pull variable (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms correctly detected the change point at sample 101 but not at sample 220 (Figure 85). This corresponds very well with the CVA biplot results (Figure 79) and the median significance results (Figure 83) where the mass pull variable could be used to more easily distinguish between class 1 and class 2 recovery data, compared to distinguishing between class 2 and class 3 recovery data. Furthermore, the SSA and ELM SSA change point detection techniques detected a very distinct and seemingly significant change point at sample 70 with the ELM SSA change point detection technique also detecting a change point at sample 140. Should the root cause of the decrease in recovery prove difficult to determine, this change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

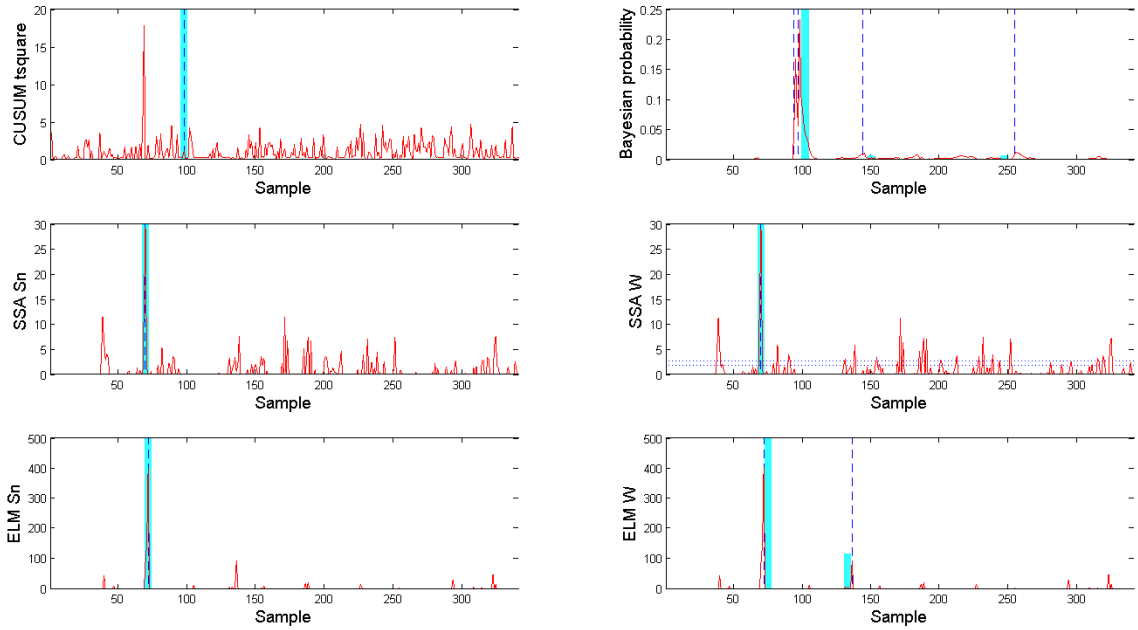

Figure 85: **Change point detection: mass pull at a confidence level of 0.99**

For the final tail grade variable (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), as for the mass pull variable, the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms correctly detected the change point at sample 101, with the Bayesian probability, the SSA and the ELM SSA change point detection algorithms correctly detecting the change point at sample 220 (Figure 86). Again, this corresponds very well with the CVA biplot results (Figure 79) and the median significance results (Figure 83) where the final tail grade variable could be used to distinguish between class 1 and class 2 recovery data as well as between class 2 and class 3 recovery data. Additionally, the Bayesian probability change point detection technique detected a small, relevant, change point at sample 270 with the SSA and ELM SSA change point detection techniques detecting a very distinct and seemingly significant change point at sample 240. As before, should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

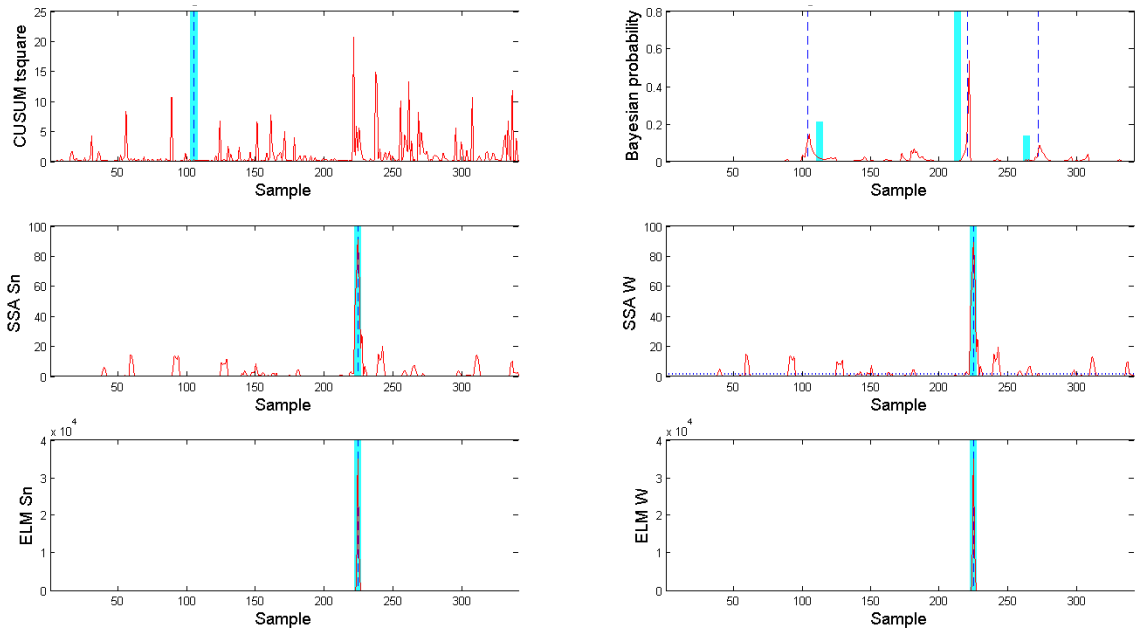

Figure 86: **Change point detection: final tail grade at a confidence level of 0.99**

For the plant feed head grade variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms correctly detected the change point at sample 101 but not at sample 220 (Figure 87). As before, this corresponds very well with the CVA biplot results (Figure 79) and the median significance results (Figure 83). Additionally, the SSA change point detection technique detected a change point at sample 200 with the ELM SSA change point detection technique detecting a change point at sample 175. As before, should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

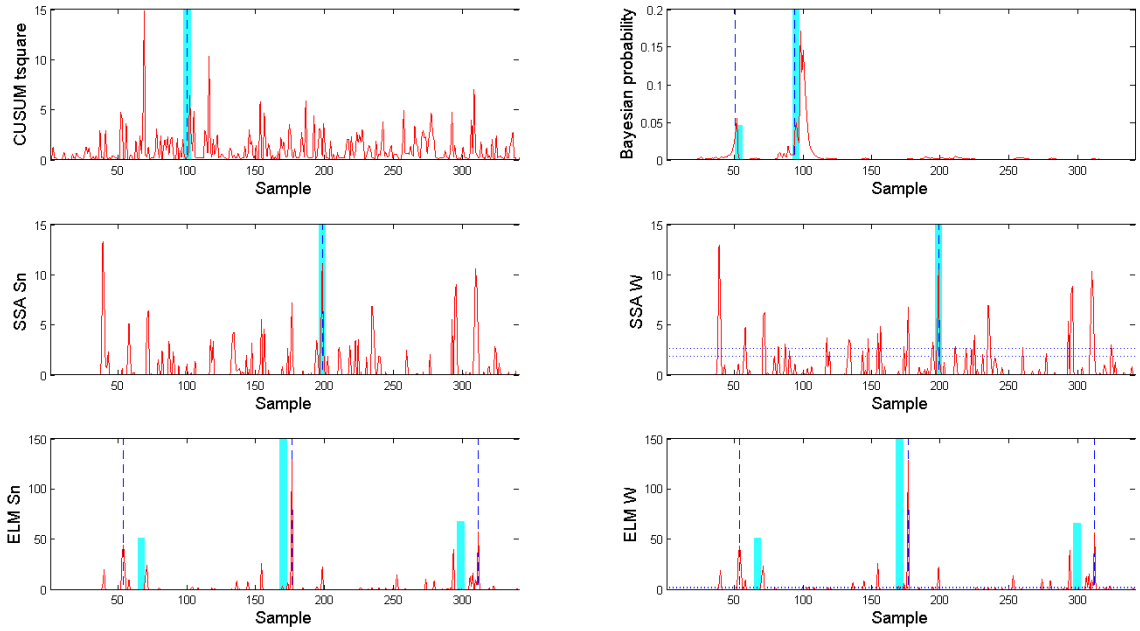

Figure 87: **Change point detection: plant feed head grade at a confidence level of 0.99**

For the final tail grind variables (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), only the Bayesian probability change point detection algorithm correctly detected the change point at sample 101, with none convincingly identifying a change point at sample 220 (Figure 88). As before, this corresponds very well with the CVA biplot results (Figure 79) and the cumulative % passing grind curve results (Figure 84). Many additional seemingly distinct and significant change points were detected, most notably at samples 50, 140 and 260 for the Bayesian change point detection algorithm, at sample 260 for the SSA change point detection algorithm and at sample 120 for the ELM SSA change point detection algorithm. It is suspected that this increase in potential change points are indicative of the fact that the grind data is more variable (noisy), frequently changing from day to day, when compared to the grade data, where less frequent shifts are observed (the data showing more of a filtered response). This is in accordance with the fundamental operation of a concentrator process where more variability exists around the operation of the comminution circuit (responsible for grind) when compared to the flotation circuit (responsible for grade). As before, should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

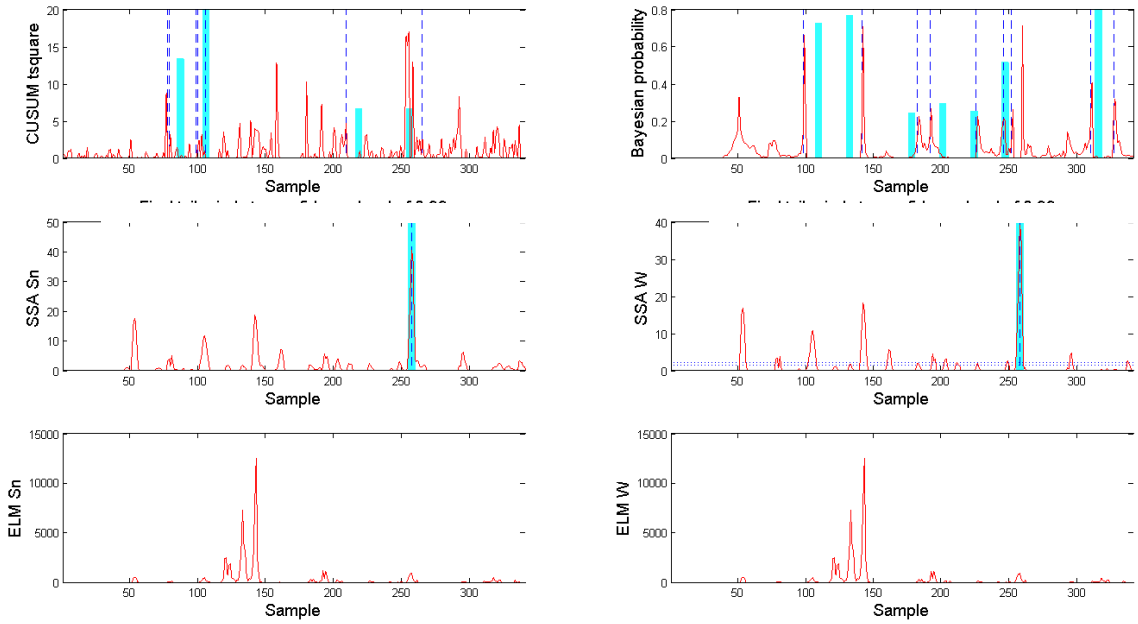

Figure 88: **Change point detection: final tail grind at a confidence level of 0.99**

At this stage of the analysis it is evident that a decrease in plant feed head grade, a decrease in mass pull, a coarser final tail grind and an increase in final tail grade all contributed to a decrease in recovery. Of these, a coarser final tail grind alludes to the fact that maybe either the silica circuit or the chrome circuit is being overloaded and not performing optimally.

As stated earlier, these contributors to the decrease in recovery now need to be further investigated. Of the contributors, the plant feed head grade measurement lies at the input side of the concentrator process and can be treated as a known disturbance root cause, not requiring further investigation at this point. The mass pull variable on the other hand, being an average contributor over all the recovery classes, should not be investigated in isolation but analysed as part of flotation performance further on. Consequently, the focus is next on identifying the drivers that have caused a shift in the final tail grind and final tail grade, resulting in a decrease in recovery.

7.1.5 Drivers: Final tail grade and grind

From a process causality map (Figure 64) perspective, the final tail grade and grind (Figure 89) is a function of the silica circuit tails grade and grind and the chrome circuit tails grade and grind.

Stellenbosch University http://scholar.sun.ac.za

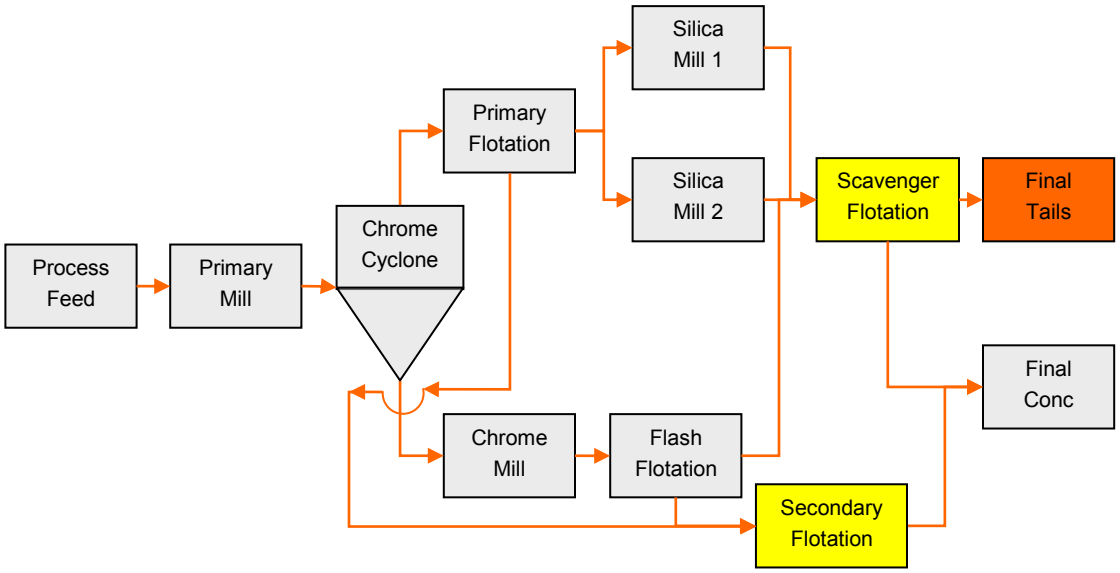

Figure 89: **Drivers: final tail**

As before, fault detection models were constructed using only the class 1 reference data from the final tail drivers data set with performance metrics calculated for the class 2 changeover and class 3 fault data. Considering this data set is only a subset of the combined data set, the univariate performance metrics delivered the exact same results (Figure 90), and therefore conclusions, as for the evaluation of the combined data set. For the reliability index (Figure 90) it was found that only five non-linear multivariate performance metrics, two dynamic multivariate performance metrics and one univariate performance metric could effectively detect the changeover and fault data. This again strongly suggests that some of the other process variables found in the combined data set are probably better indicators of a fault condition in the process compared to the process variables in the final tail drivers data set. Alternatively, the possibility also exist that the magnitude of the fault condition as represented by the selected process variables was too small to result in a measurable process performance degradation over the evaluation period (variation still within variation associated with common cause).

Stellenbosch University http://scholar.sun.ac.za

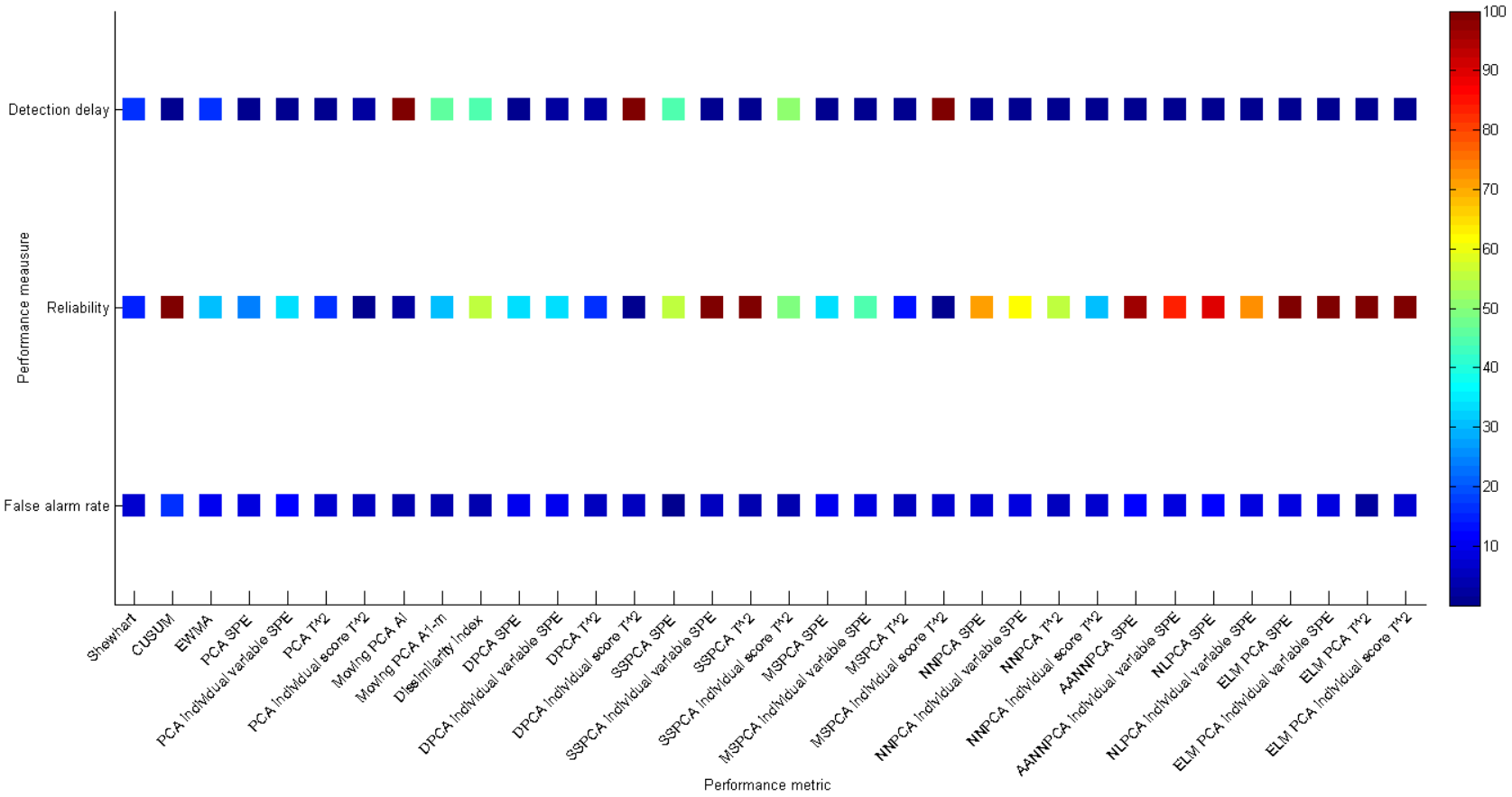

Figure 90: **Statistical data-based fault detection: final tail grade and grind – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 91) is made to determine if the final tail grade and grind drivers can be used to distinguish between the different recovery classes.

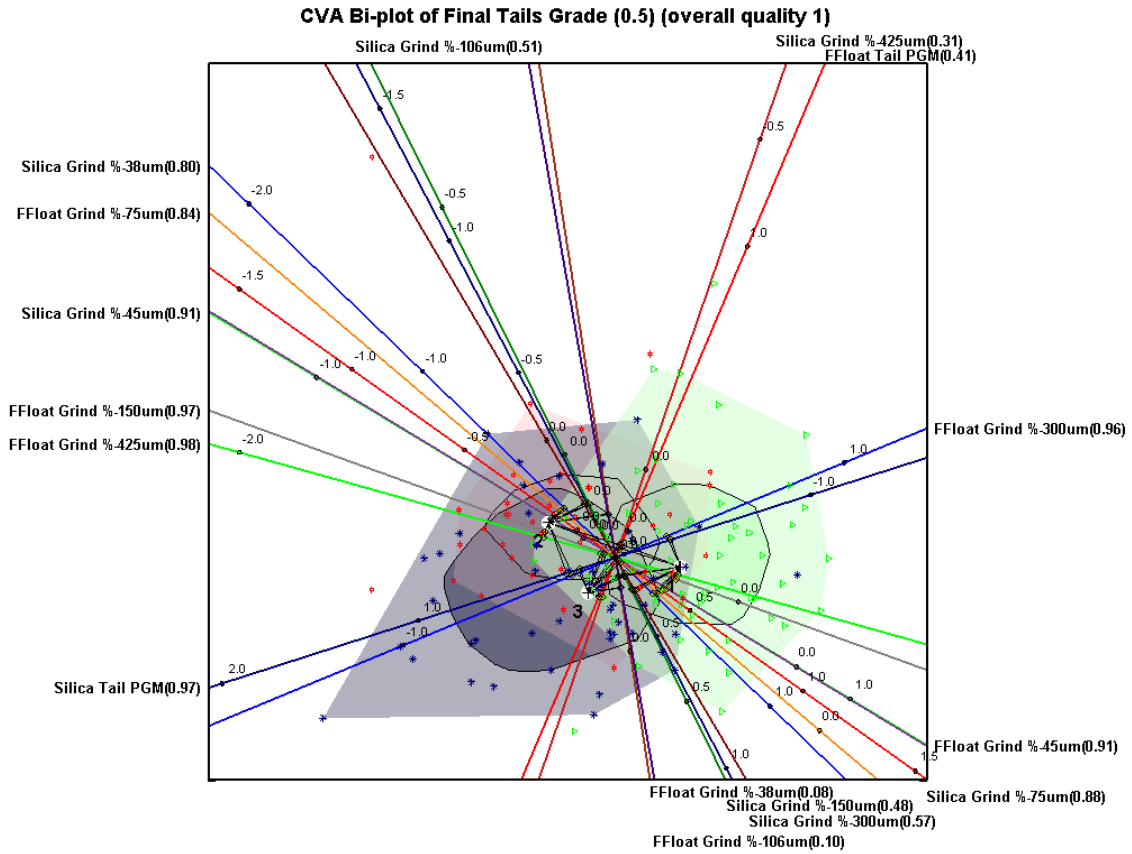

Figure 91: **CVA biplot: final tail grade and grind (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

From the CVA biplot it is evident that the classes are not well separated with the alpha bags overlapping each other even at an alpha value of 0.5. With the chrome circuit tails grade (FFloat Tail PGM) variable axis and most of the silica circuit tails grind (Silica Grind) variable axes having relatively low predictivities, significant and reliable conclusions can only be drawn with regards to the discriminatory ability of the silica circuit tails grade (Silica Tail PGM) and the chrome circuit tails grind (FFloat Grind).

There seems to exist a high degree of correlation between the silica circuit tails grind and the chrome circuit tails grind (both either getting coarser or finer at the same time), with the silica circuit tails grade and chrome circuit tails grade being more correlated to each other than to the tails grind (as the silica circuit tails grade decreases, the chrome circuit tails grade increases). Fundamentally this combination does not make sense, and since the silica circuit tails grind variable axes have relatively low predictivities, its correlation to the chrome circuit tails grind is probably weak at best or due to the process being

Stellenbosch University http://scholar.sun.ac.za

operated in some abnormal state. Visually, the CVA biplot shows that on average over all 3 classes the main differentiators between the recovery classes are the silica circuit tails grade and the chrome circuit tails grind (in the form of the flash float grind %-300 µm). When moving only from class 1 to class 2, these drivers are overshadowed by the silica circuit tails grind and the chrome circuit tails grind. Likewise, these drivers are also overshadowed when only moving from class 2 to class 3, however, in this instance by the silica circuit tails grind %-425 µm and the chrome circuit tails grade (however, both having relatively low axes predictivities, resulting in conclusions drawn about these variables being unreliable). From this visual inspection and analysis of the CVA biplot results it would seem as if the biggest shift occurred in the silica circuit, with a smaller, less prominent shift occurring in the chrome circuit.

As before, the CVA biplot analysis is supported by the variable importance analysis. From the variable importance analysis on average over all 3 classes (Figure 92) the main differentiators between the recovery classes are the fine silica circuit tails grind component and the coarse chrome circuit tails grind component. Moving only from class 1 to class 2 (Figure 93), the fine silica circuit tails grind component and the coarse chrome circuit tails grind component again stood out as important drivers with the fine chrome circuit tails grind component joining these drivers as important when moving from class 2 to class 3 (Figure 94). These results correspond very well with the findings from analysing the CVA biplot. Consequently it is confirmed that a bigger shift in the process operation of the silica circuit occurred when compared to the process operation of the chrome circuit. For completeness, both the grade and grind of both the silica and chrome circuits will be analysed.

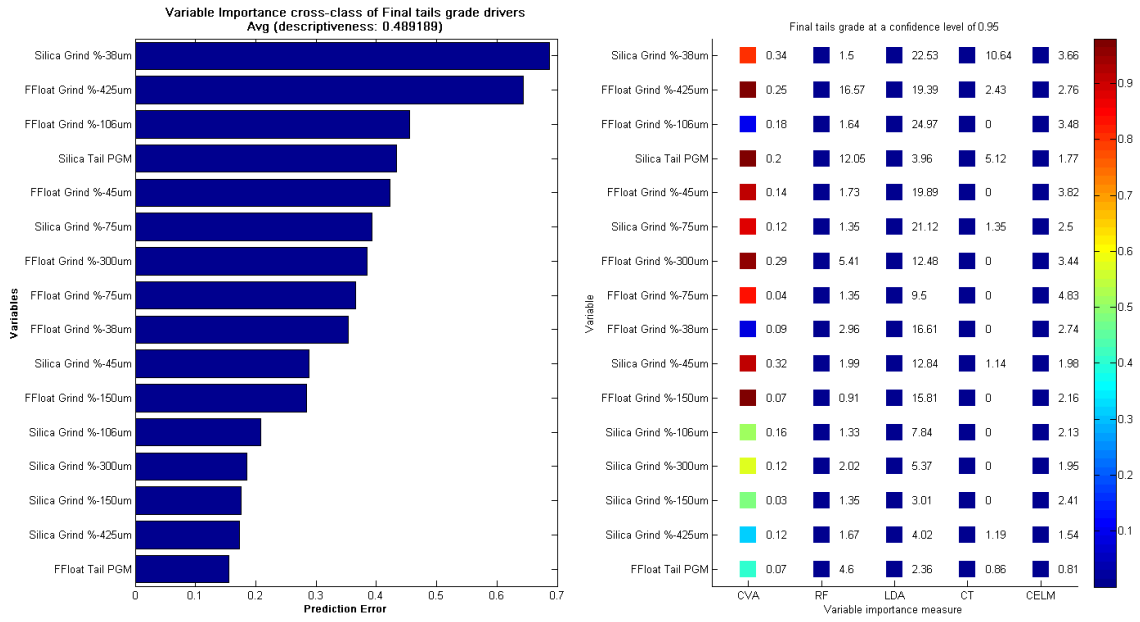

Figure 92: **Variable importance class 1-2-3: final tail grade and grind at a confidence level of 0.95**

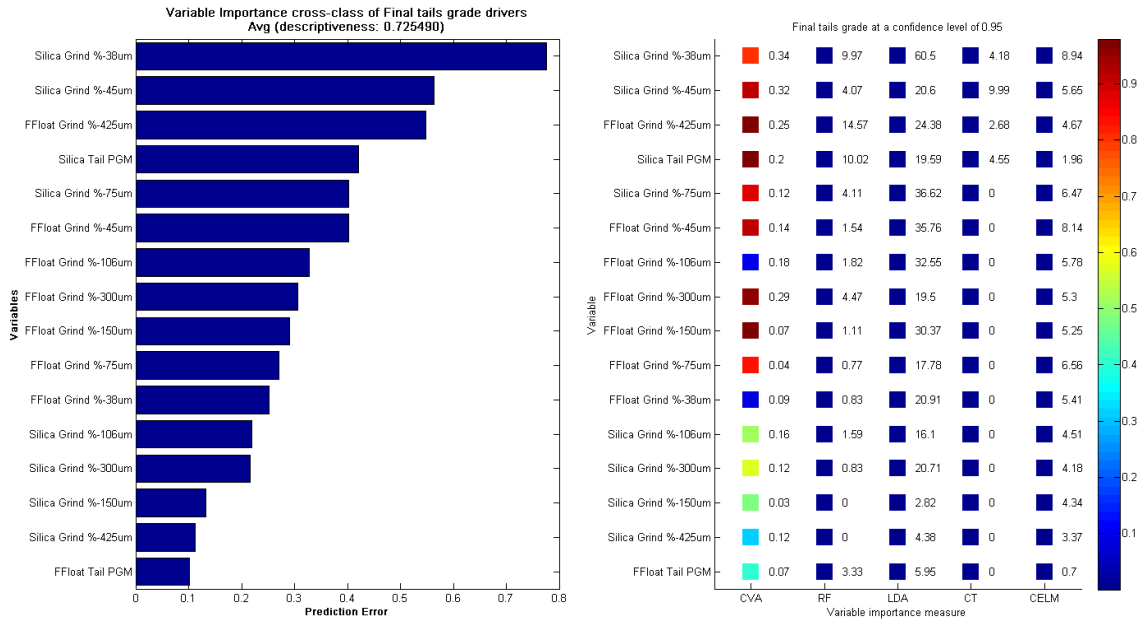

Figure 93: **Variable importance class 1-2: final tail grade and grind at a confidence level of 0.95**

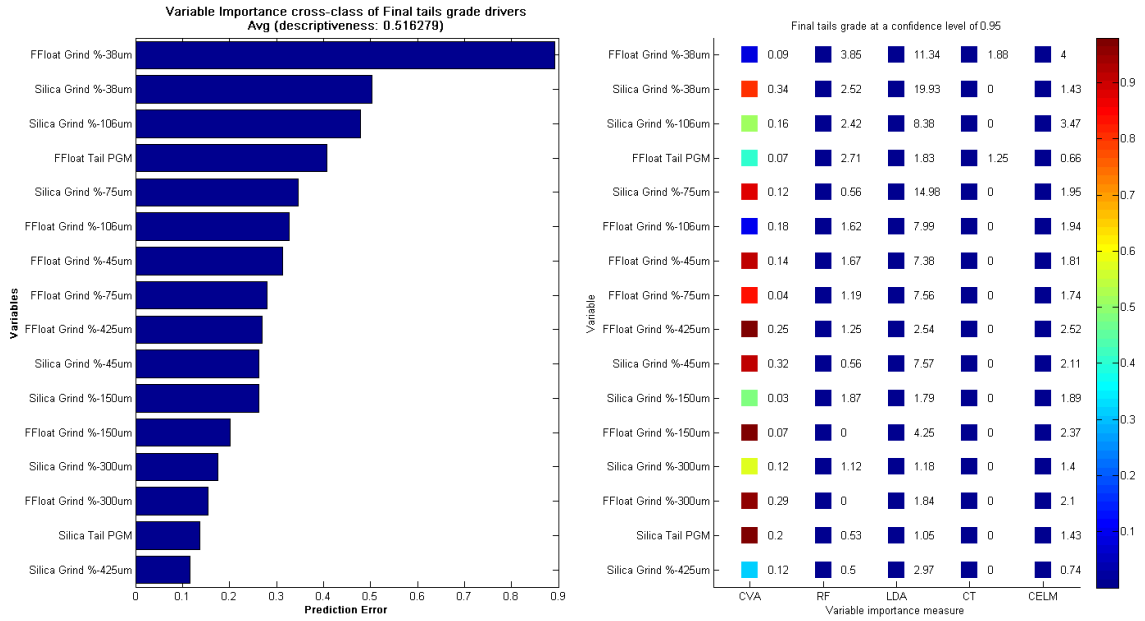

Figure 94: **Variable importance class 2-3: final tail grade and grind at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

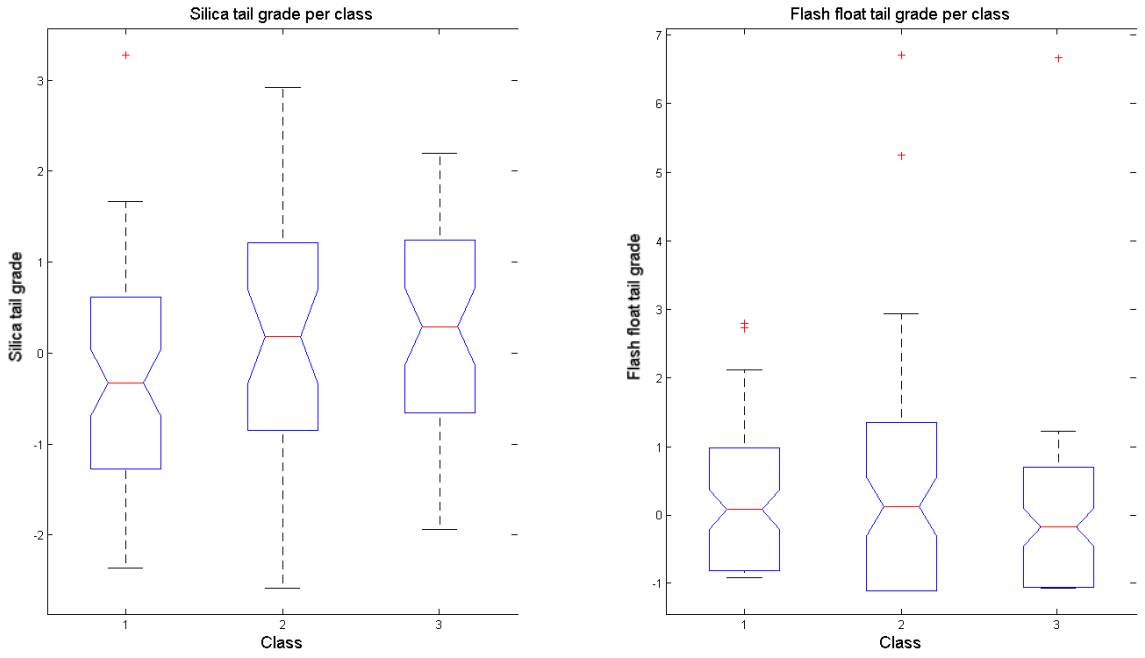

Figure 95: **Median significance: silica and chrome circuit tails grade**

Subjecting the 3 silica circuit tails grade and chrome circuit tails grade classes of data to the median significance (Figure 95) and one-way ANOVA analysis, it was found that there was only a significant difference between the classes of the silica circuit tails grade at a significance level of less than 0.05. For the grind comparison (Table 13), it was found that only the smaller silica circuit tails grind fractions (%-38 µm, %-45 µm and %-75 µm) and the larger chrome circuit tails grind fractions (%-300 µm and %-425 µm) showed a significant difference at a significance level of less than 0.05 (matching the results from the variable importance analysis for class 1-2-3), with the silica circuit tails grind becoming noticeably coarser and the chrome circuit tails grind almost not being affected at all (Figure 96).

These findings support the notion of a bigger shift having occurred in the silica circuit, as is evident from an increased silica circuit tails grade and a coarser silica circuit tails grind which would have resulted in an increase final tails grade, when compared to the chrome circuit. There does, however, seem to be a certain amount of mismatch between the initial recovery classes and the subsequent silica and chrome circuit tails grade and grind classes. Should the root cause of the decrease in recovery prove difficult to determine, the data should be re-classed into more distinct silica and chrome circuit tails grade and grind classes that still aligns with the recovery classes.

Stellenbosch University http://scholar.sun.ac.za

Table 13: **Grind significance: silica and chrome circuit tails grind**

Variable	Significance value
Silica Grind %-38 µm	0.0368
Silica Grind %-45 µm	0.0332
Silica Grind %-75 µm	0.0307
Silica Grind %-106 µm	0.3666
Silica Grind %-150 µm	0.7397
Silica Grind %-300 µm	0.9625
Silica Grind %-425 µm	0.8962
Flash Float Grind %-38 µm	0.8141
Flash Float Grind %-45 µm	0.9404
Flash Float Grind %-75 µm	0.8890
Flash Float Grind %-106 µm	0.8014
Flash Float Grind %-150 µm	0.9847
Flash Float Grind %-300 µm	0.0018
Flash Float Grind %-425 µm	3.458e-05

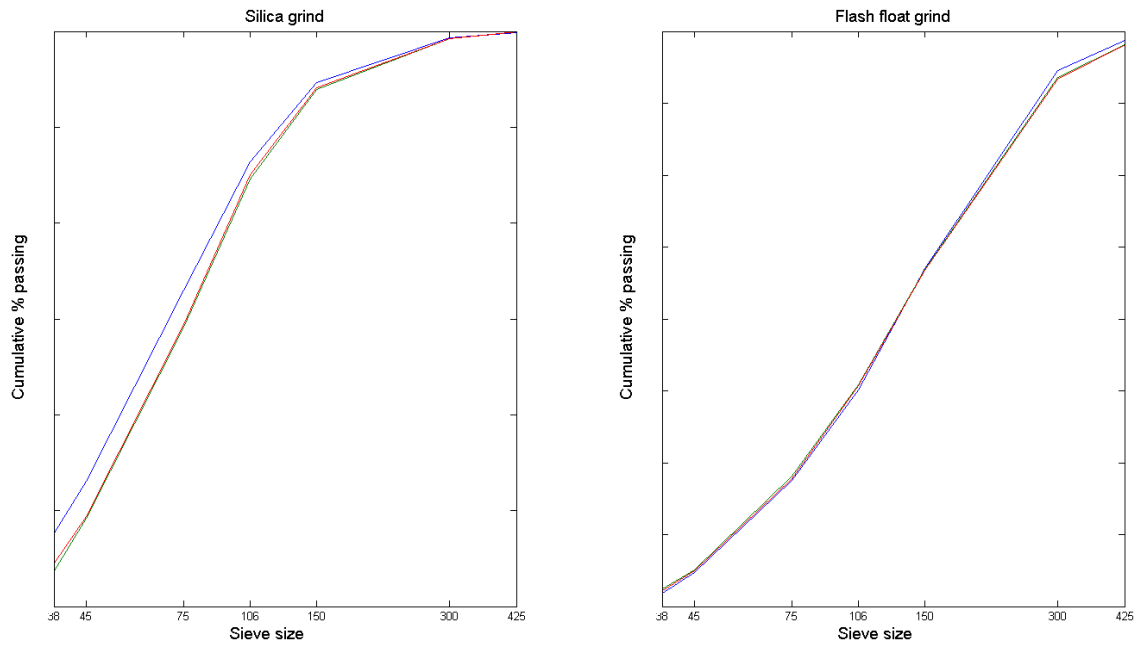

Figure 96: **Cumulative % passing grind curve: silica and chrome circuit tails grind (class 1 – blue; class 2 – green; class 3 – red)**

Although a good correlation has been found thus far between the change points detected in the recovery variable and potential important variables, especially related to the silica circuit, it would be interesting to

Stellenbosch University http://scholar.sun.ac.za

see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others.

For the silica circuit tails grade variable (not autocorrelated, not normally distributed, not exponentially distributed, Bayesian suggested change point detection technique), the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms correctly detected the change point at sample 101 but not at sample 220 (Figure 97). This corresponds very well with the CVA biplot results (Figure 91) and the median significance results (Figure 95) where the silica circuit tails grade could be used to more easily distinguish between class 1 and class 2 recovery data, compared to distinguishing between class 2 and class 3 recovery data. Many additional seemingly distinct and significant change points were also detected, most notably at sample 250, contributing to the likely requirement that the data should be re-classed. Should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

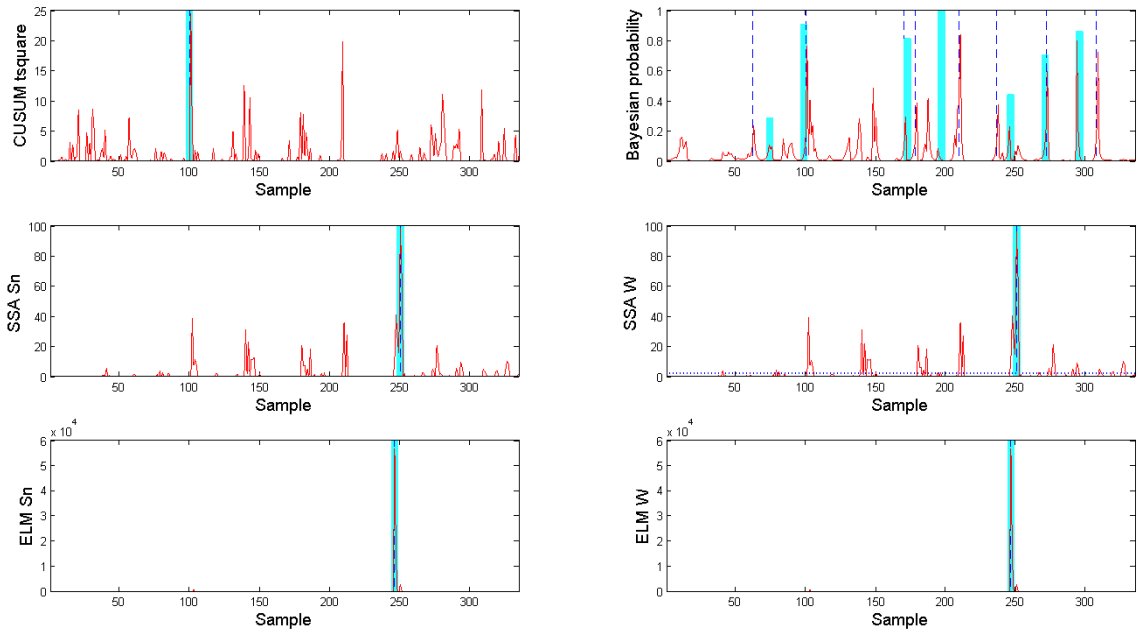

Figure 97: **Change point detection: silica circuit tails grade at a confidence level of 0.99**

For the chrome circuit tails grade variable (not autocorrelated, not normally distributed, not exponentially distributed, Bayesian suggested change point detection technique), not one of the change point detection algorithms correctly detected the change points at samples 101 or at sample 220 (Figure 98). This corresponds very well with the CVA biplot results (Figure 91) and the median significance results (Figure 95) where the chrome circuit tails grade was not significantly different for the 3 recovery classes. This clearly supports the notion that changes in the chrome circuit tails grade do not coincide with changes in

Stellenbosch University http://scholar.sun.ac.za

the recovery and therefore do not contribute significantly to the decrease in recovery. Many additional seemingly distinct and significant change points were, however, detected, most notably at samples 187 and 295, both being spikes in chrome circuit tails grade time series. Should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

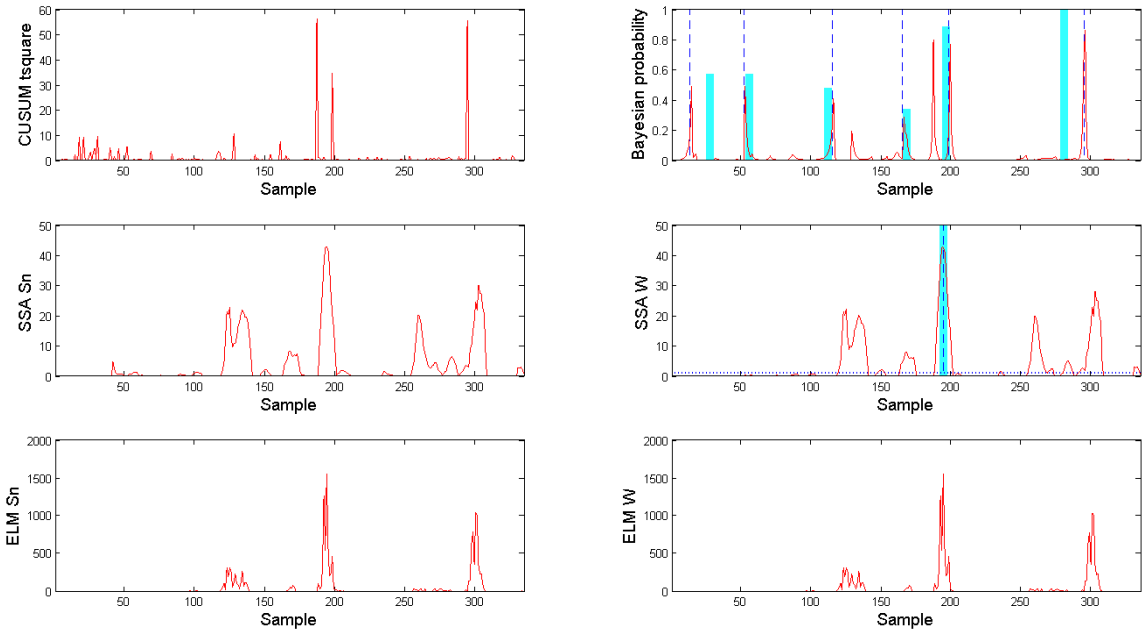

Figure 98: **Change point detection: chrome circuit tails grade at a confidence level of 0.99**

For the silica circuit tails grind variables (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), there is lots of evidence of distinct events at around sample 95 and sample 250 (Figure 99). Although these change points don't match the recovery data change points exactly, they are in relative close proximity and could be indicative of the fact that re-classing of the data at this level of the analysis is required. As for the silica circuit tails grade, this corresponds very well with the CVA biplot results (Figure 91) and the median significance results (Figure 95) where the silica circuit tails grind could be used to more easily distinguish between class 1 and class 2 recovery data, compared to distinguishing between class 2 and class 3 recovery data. As with the final tails grind, the Bayesian change point detection algorithm detected many additional seemingly distinct and significant change points. This again highlights the possibility that the increase in potential change points are indicative of the fact that the grind data is more variable, frequently changing from day to day, when compared to the grade data, where less frequent shifts are observed. As before, should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

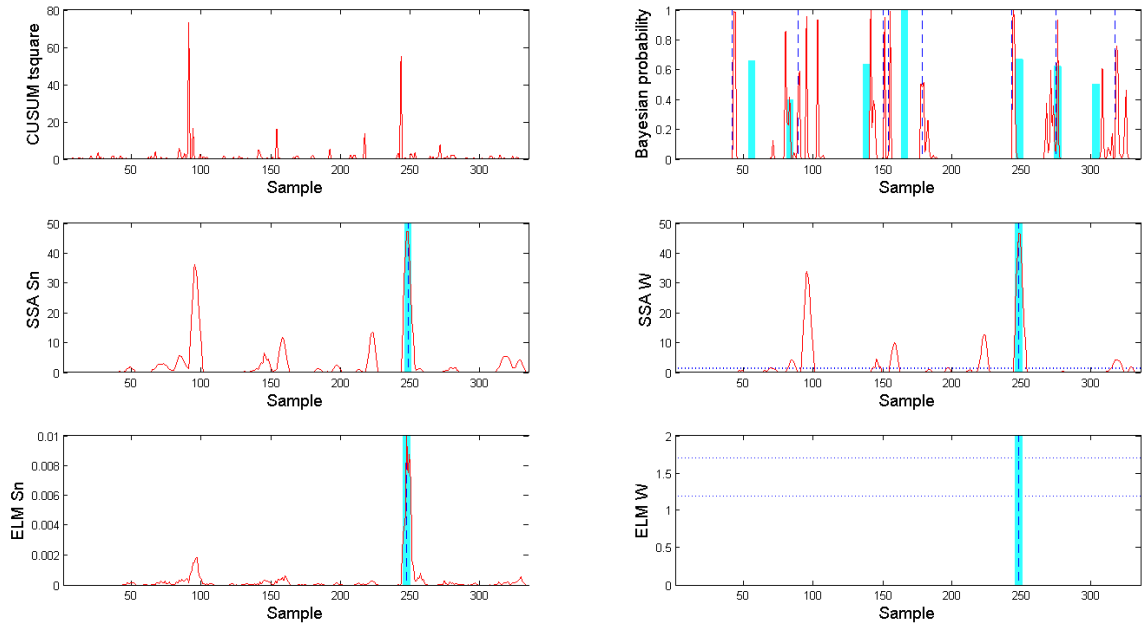

Figure 99: **Change point detection: silica circuit tails grind at a confidence level of 0.99**

For the chrome circuit tails grind variables (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), the Bayesian probability change point detection algorithm correctly detected both the change points at approximately sample 101 and sample 220 (Figure 100). However, many other seemingly significant change points were also detected; most notably at approximately sample 175 (by most of the change point detection techniques) which coincides with a spike in the chrome circuit tails grind time series data. From previous findings it can be concluded that the increase in potential change points detected by the Bayesian probability algorithm, including those at approximately sample 101 and sample 220, is in all likelihood due to an increase in day to day variation in the grind data and not necessarily due to relevant events in the data. As before, however, should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

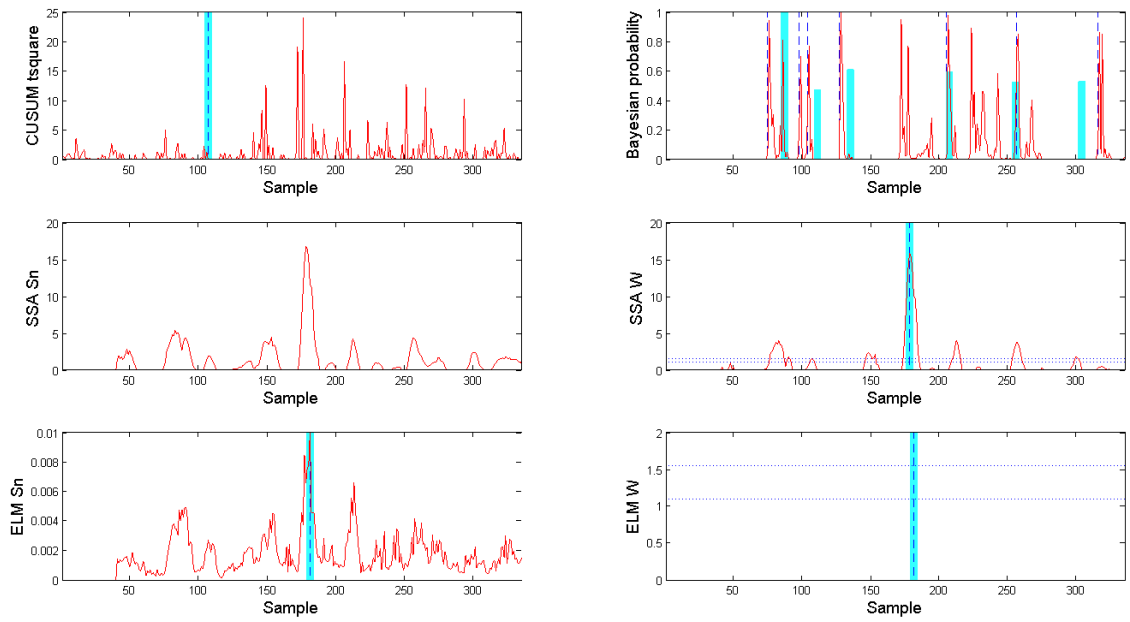

Figure 100: **Change point detection: chrome circuit tails grind at a confidence level of 0.99**

At this stage of the analysis it is evident that primarily an increase in the silica circuit tails grade and a coarser silica circuit tails grind, especially considering the shift from the recovery class 1 to class 2 data, contributed to a shift in the final tails grade and grind which in turn contributed to a decrease in the recovery. Unfortunately, no reliable data is available regarding the silica circuit flotation process and the assumption has to be made that the increase in the silica circuit tails grade is due to the coarsening of the silica circuit grind. A shift in the chrome circuit performance was much less prominent than the shift in the silica circuit performance, therefore, being less important when considering the decrease in the recovery, resulting in less attention being given to investigating it. Consequently, the focus is next on identifying the drivers that have caused a shift in the silica circuit tail grind, ultimately resulting in a decrease in the recovery.

7.1.6 Drivers: Silica circuit grind

From a process causality map (Figure 64) perspective, the silica circuit tails grind is a function of the silica circuit milling (Figure 101). The silica circuit milling in turn is a function of variables such as mill power, mill load, mill throughput and mill feed particle size distribution, in the form of the primary flotation (rougher) tail grind PSD in this instance (Figure 60).

Stellenbosch University http://scholar.sun.ac.za

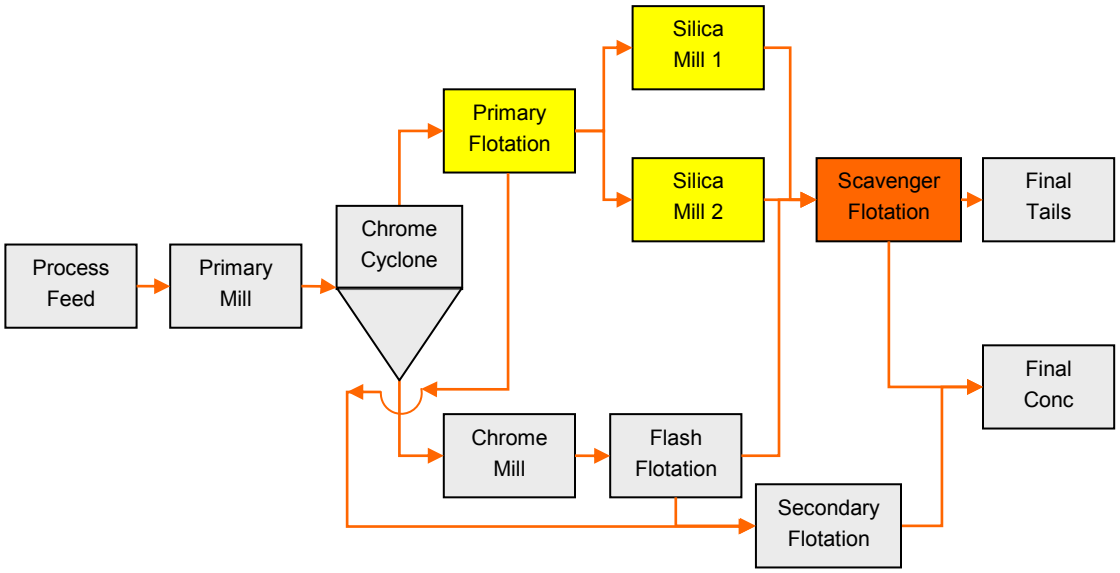

Figure 101: **Drivers: silica circuit**

As before, fault detection models were constructed using only the class 1 reference data from the silica circuit drivers data set with performance metrics calculated for the class 2 changeover and class 3 fault data. Considering this data set is only a subset of the combined data set, the univariate performance metrics delivered the exact same results (Figure 102), and therefore conclusions, as for the evaluation of the combined data set. For the reliability index (Figure 102) it was found that the non-linear multivariate performance metrics performed the best, followed by the basic multivariate performance metrics and the univariate performance metrics with none of the dynamic multivariate performance metrics reliably detecting the fault condition. When compared to the fault detection results for the final tail drivers data set, it is evident by the increased number of reliable performance metrics that the fault condition is more prominent in the silica circuit drivers data set. This can potentially be ascribed to the fact that the magnitude of the fault condition as represented by the selected process variables in the silica circuit drivers data set is large enough to result in a measurable process performance degradation over the evaluation period (variation exceeding variation associated with common cause).

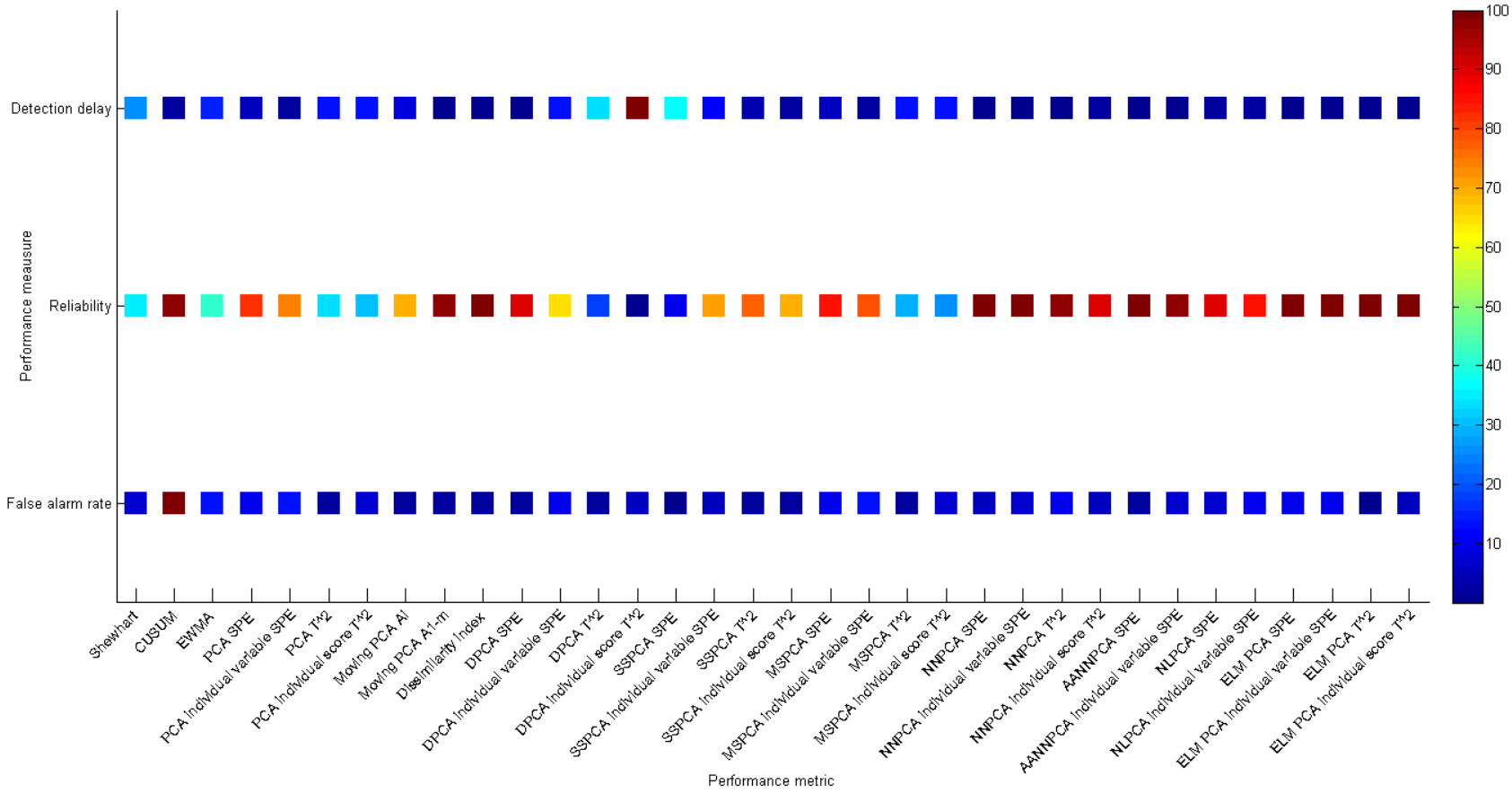

Figure 102: **Statistical data-based fault detection: silica circuit grind – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 103) is made to determine if the silica circuit grind drivers can be used to distinguish between the different recovery classes.

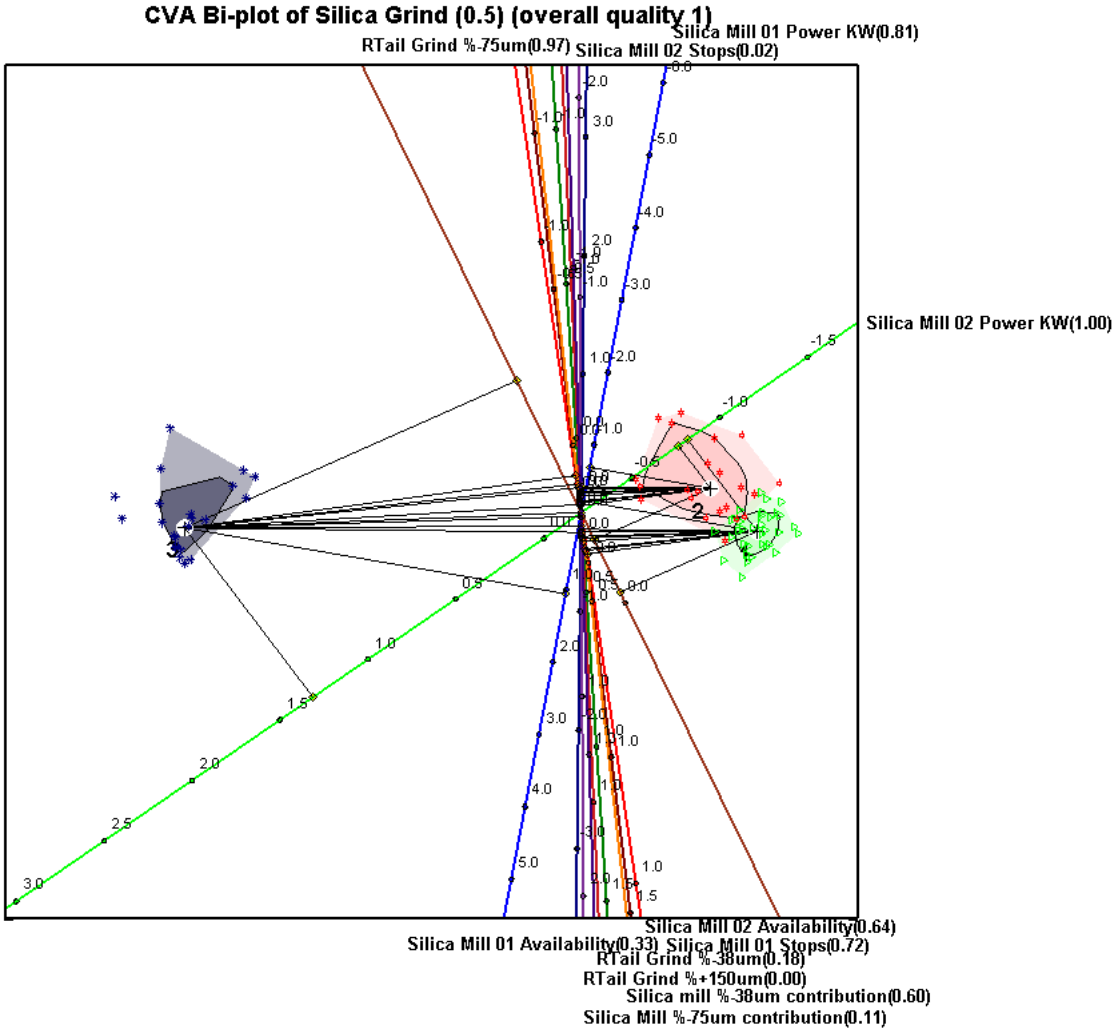

Figure 103: **CVA biplot: silica circuit grind (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

From the CVA biplot it is evident that the data based on the recovery classes are very well separated, with only slight overlap between class 1 and class 2 for the alpha bags with a significance of 50%. Regarding the axes predictivities, the silica mill 01 power, silica mill 02 power and primary flotation (RTail) grind %-75 µm, have the highest, and therefore most reliable scores, allowing significant and reliable conclusions to be drawn regarding their discriminatory ability when considering the silica circuit grind data.

Visually, the CVA biplot shows that class 1 and class 2 data is closely related to each other, both which in turn is significantly removed from class 3 data. On average, over all 3 classes the main differentiator

Stellenbosch University http://scholar.sun.ac.za

between the recovery classes is the primary flotation grind %-75 µm variable, with the silica mill 01 power also seemingly significant. These variables are also the main differentiators between classes when moving from class 1 to class 2. However, when moving from class 2 to class 3, these drivers are overshadowed by the silica mill 02 power variable. Fundamentally, considering the results from the visual inspection of the CVA biplot analysis, it would seem as if a shift in the chrome classification performance occurred.

As before, the CVA biplot analysis is supported by the variable importance analysis. From the variable importance analysis on average over all 3 classes (Figure 104) the main differentiators between the recovery classes are the silica mill 01 and 02 power, and to a much lesser extent, the primary flotation grind. Moving only from class 1 to class 2 (Figure 105), the silica mill 01 and 02 power variables are still the most important, however, the primary flotation grind has become a more prominent driver. Subsequently, moving from class 2 to class 3 (Figure 106), as with the CVA biplot analysis, the silica mill 02 power variable is the most significant driver, with all the other variables considered insignificant. These results correspond very well with the findings from analysing the CVA biplot, requiring the silica mill 01 power, silica mill 02 power and primary flotation grind to be further analysed.

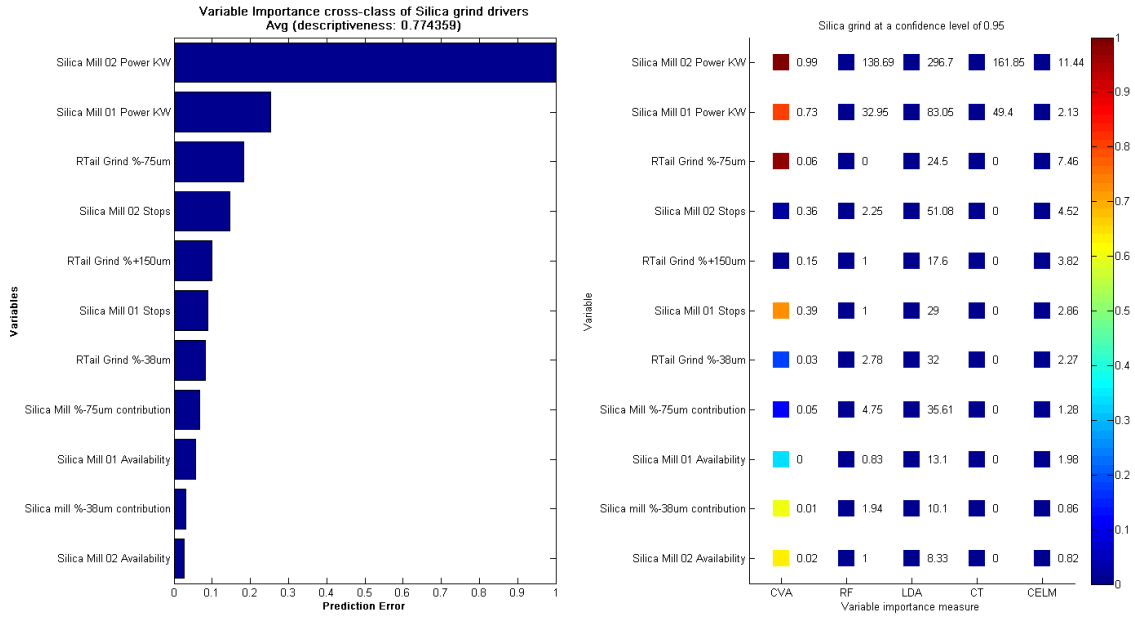

Figure 104: **Variable importance class 1-2-3: silica circuit grind at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

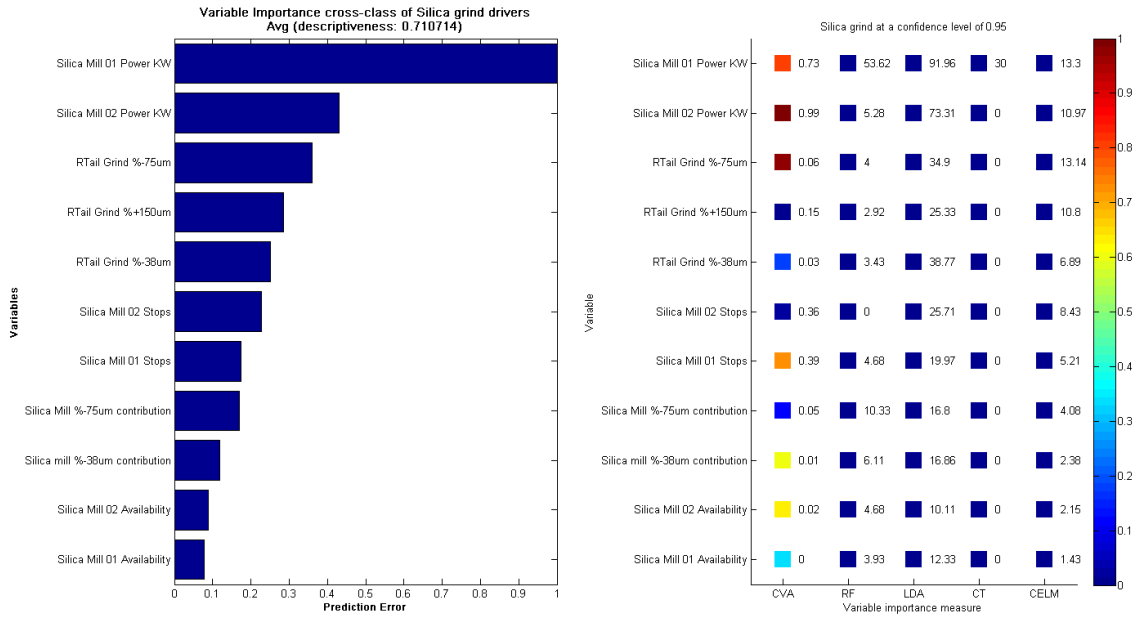

Figure 105: **Variable importance class 1-2: silica circuit grind at a confidence level of 0.95**

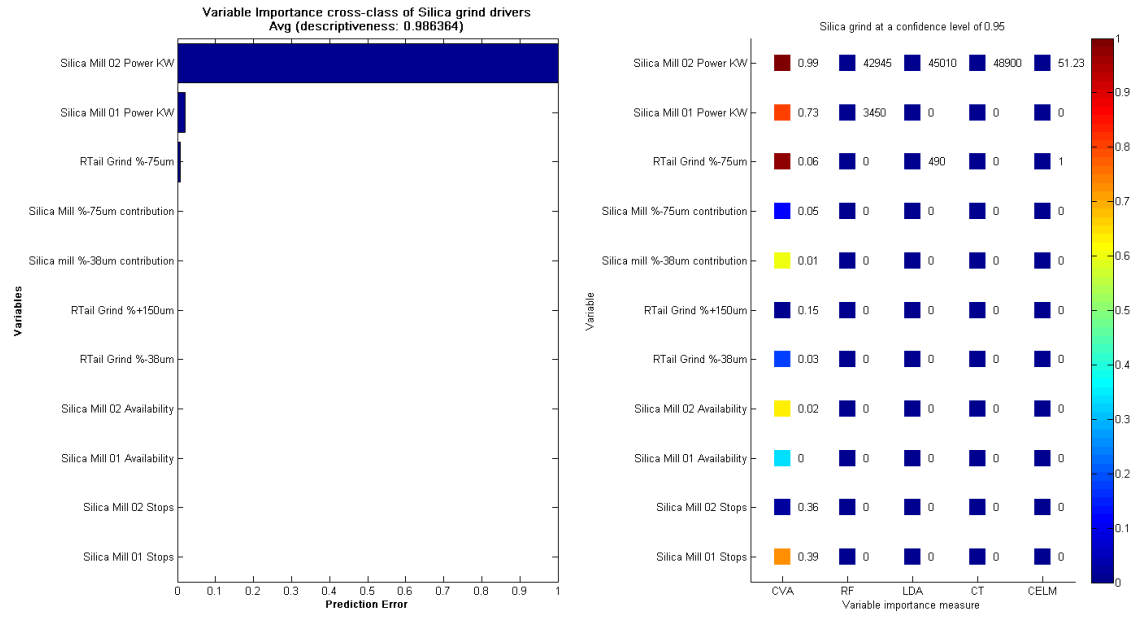

Figure 106: **Variable importance class 2-3: silica circuit grind at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

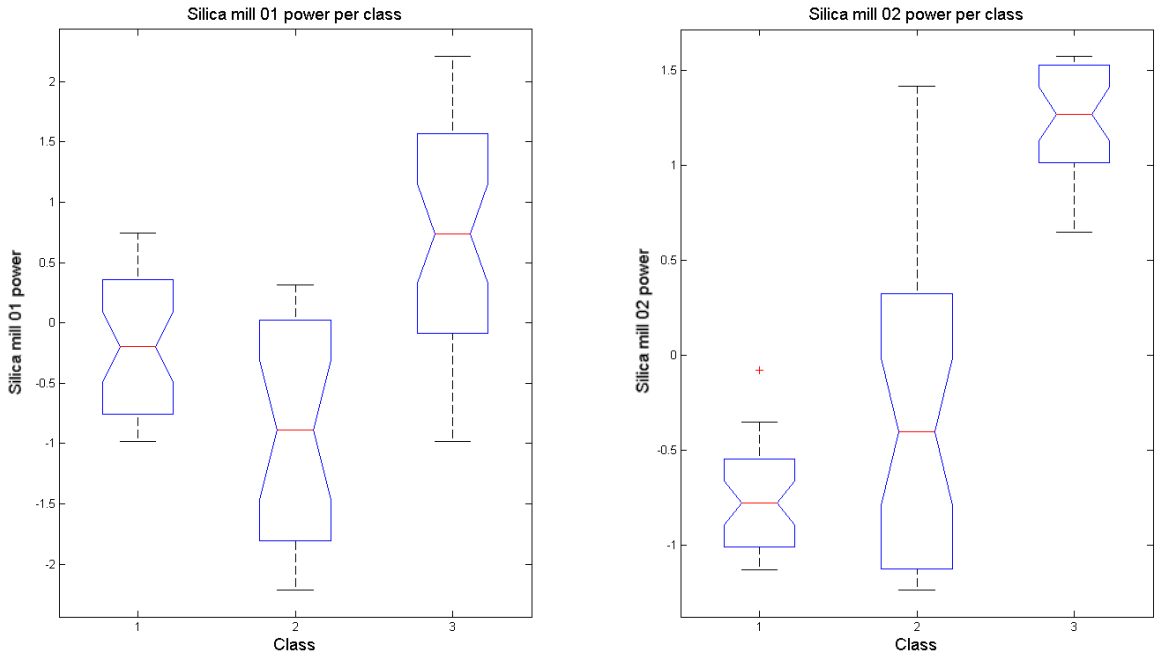

Figure 107: **Median significance: silica mill 01 and 02 power**

Subjecting the 3 silica mill 01 power and silica mill 02 power classes of data to the median significance (Figure 107) and one-way ANOVA analysis, it was found that there was a significant difference between the classes for both these variables at a significance level of less than 0.05. For the grind comparison (Table 14), it was found that there was not a significant difference between the classes for any of the particle size fraction at a significance level of less than 0.05. Visually, however, there does seem to have been a shift in the primary flotation grind with the grind initially becoming finer and then coarsening again slightly. As with the CVA biplot and the variable importance analysis, this would seem to confirm that a shift has occurred in the chrome classification performance.

Table 14: **Grind significance: primary flotation tail grind**

Variable	Significance value
Rougher Tail Grind %-38 µm	0.9061
Rougher Tail Grind %-75 µm	0.8663
Rougher Tail Grind %+150 µm	0.2950

Stellenbosch University http://scholar.sun.ac.za

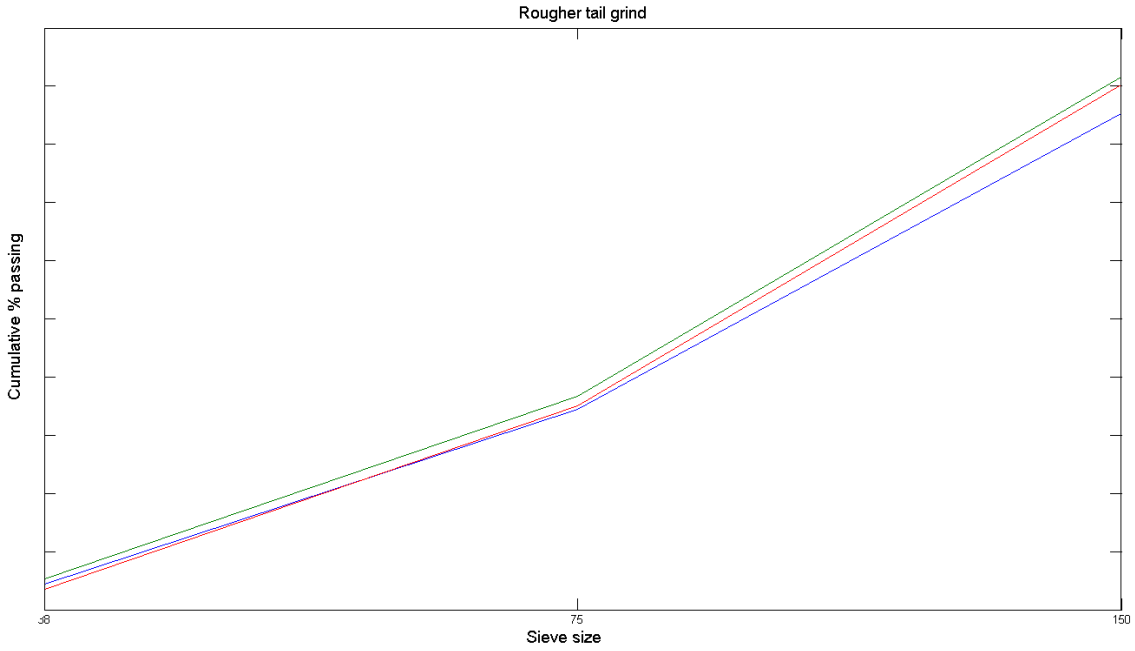

Figure 108: **Cumulative % passing grind curve: primary flotation tail grind (class 1 – blue; class 2 – green; class 3 – red)**

Although a good correlation has been found thus far between the change points detected in the recovery variable and potential important variables, it would again be interesting to see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others. Whereas the initial data was sampled shiftly (8 hourly), the silica circuit data was sampled daily (24 hourly). Subsequently, the comparative recovery change points for this data set can be found at sample 40 (08/11/2007) and 75 (13/12/2007). Since the change in the primary flotation tail grind, based on the recovery classes, did not prove to be significant, the primary rougher flotation grind will not be subjected to change point detection and the primary flotation circuit performance not analysed.

For the silica mill 01 power variable (autocorrelated, normally distributed, not exponentially distributed, SSA suggested change point detection technique), the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms detected a change at sample 50 (10 day subsequent to the first recovery change point), returning to its prior state at sample 65 (as indicated by the nearest-neighbours CUSUM T^2 plot), and at sample 75 (coinciding with the second recovery change point), also returning to its prior state at sample 95 (Figure 109). This corresponds well with the CVA biplot results (Figure 103), where the silica mill 01 power was not a perfect differentiator between the different classes, and the median significance results (Figure 107) where there was still overlap between the data when

Stellenbosch University http://scholar.sun.ac.za

considering the silica mill 01 power classes. Few other significant change points, most notably at sample 80, were detected and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

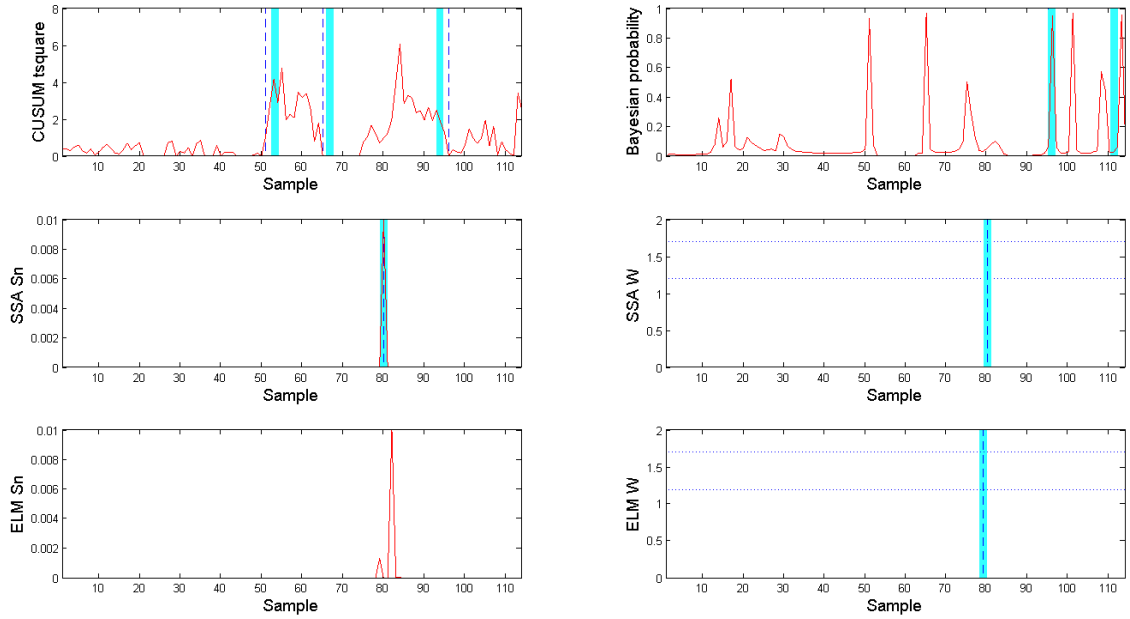

Figure 109: **Change point detection: silica mill 01 power at a confidence level of 0.99**

For the silica mill 02 power variable (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), the nearest-neighbours CUSUM and Bayesian probability change point detection algorithms did not detect the first recovery change point, but did detect a change at sample 65 (10 day prior to the second recovery change point), seemingly shifting to another state at approximately sample 95, as indicated by the nearest-neighbours CUSUM T^2 plot (Figure 110). Again, these results correspond very well with the CVA biplot results (Figure 103), where the silica mill 02 power was not a good differentiator between the recovery class 1 and class 2 but an excellent differentiator between the recovery class 2 and class 3, and the median significance results (Figure 107) where there was still overlap between the class 1 and class 2 data, but significant separation between these and the class 3 data.

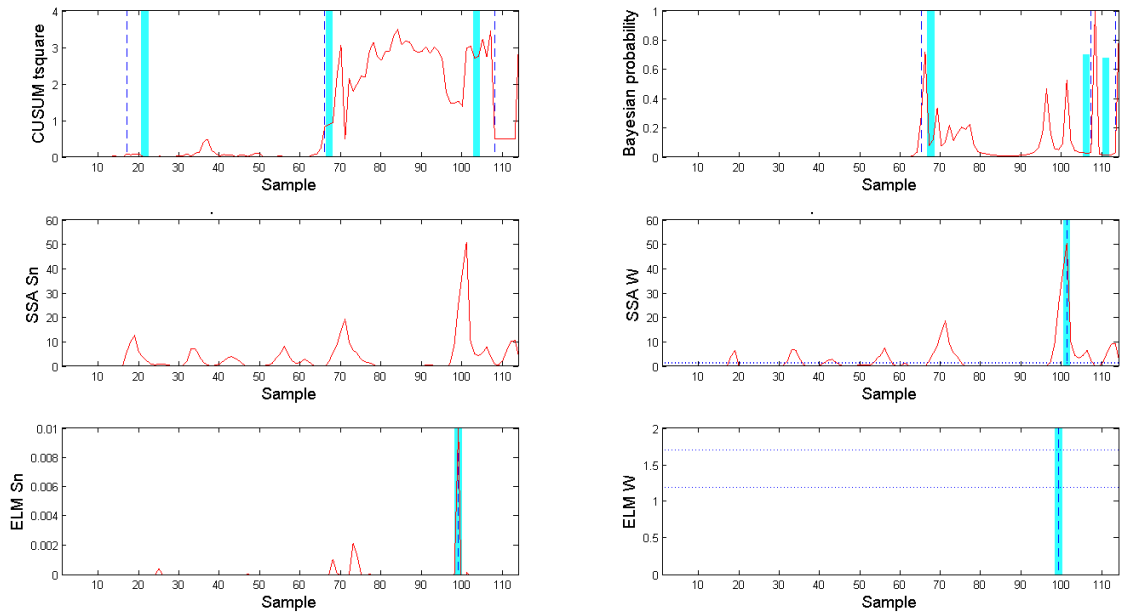

Figure 110: **Change point detection: silica mill 02 power at a confidence level of 0.99**

The significance of the changes in the silica mills, especially the silica mill power, can be better visualised using higher frequency data. OPM graphs, in the form of distribution plots, are used to not only show the operating position of the variable, but also give an indication as to the stability of the variable.

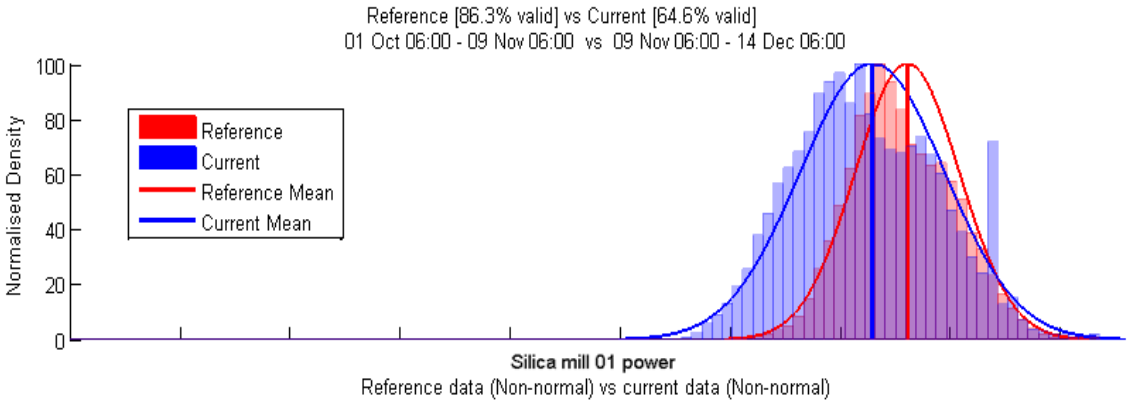

Figure 111: **OPM report: distribution comparison of silica mill 01 power (reference = class 1; current = class 2)**

Stellenbosch University http://scholar.sun.ac.za

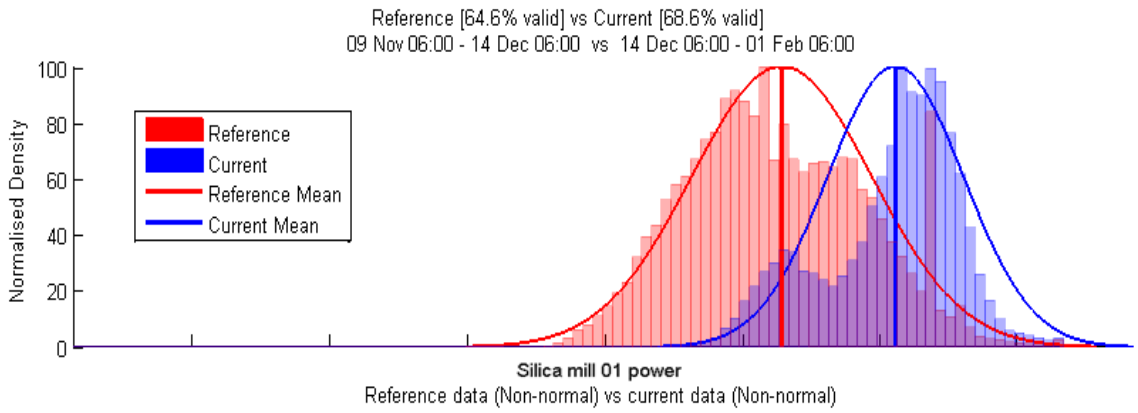

Figure 112: **OPM report: distribution comparison of silica mill 01 power (reference = class 2; current = class 3)**

From the silica mill 01 distribution plot it can be confirmed that there was a slight decrease in mill power when moving from class 1 to class 2 of the data (Figure 111) and a more significant increase in mill power when moving from class 2 to class 3 of the data (Figure 112). Similarly, from the silica mill 02 distribution plot it can be confirmed that there was a slight increase in mill power when moving from class 1 to class 2 of the data (Figure 113) and a significant increase in mill power when moving from class 2 to class 3 of the data (Figure 114).

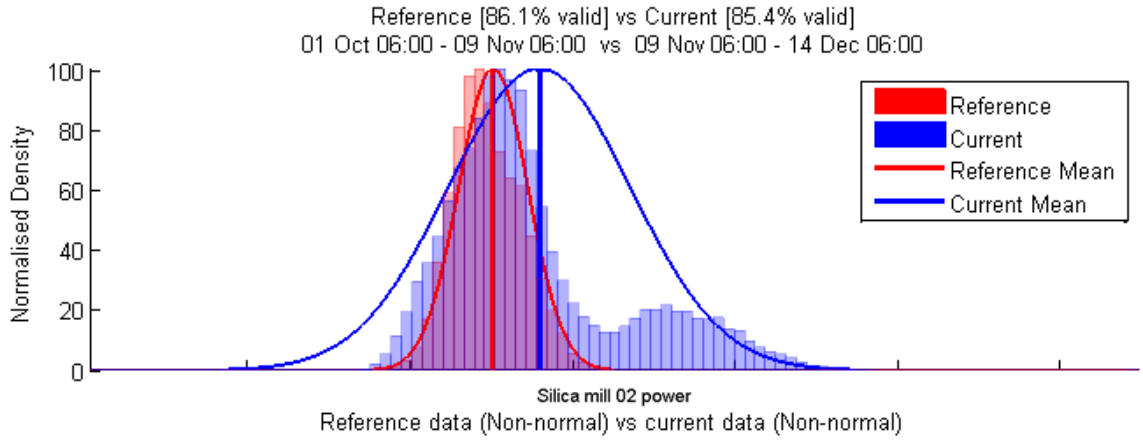

Figure 113: **OPM report: distribution comparison of silica mill 02 power (reference = class 1; current = class 2)**

Stellenbosch University http://scholar.sun.ac.za

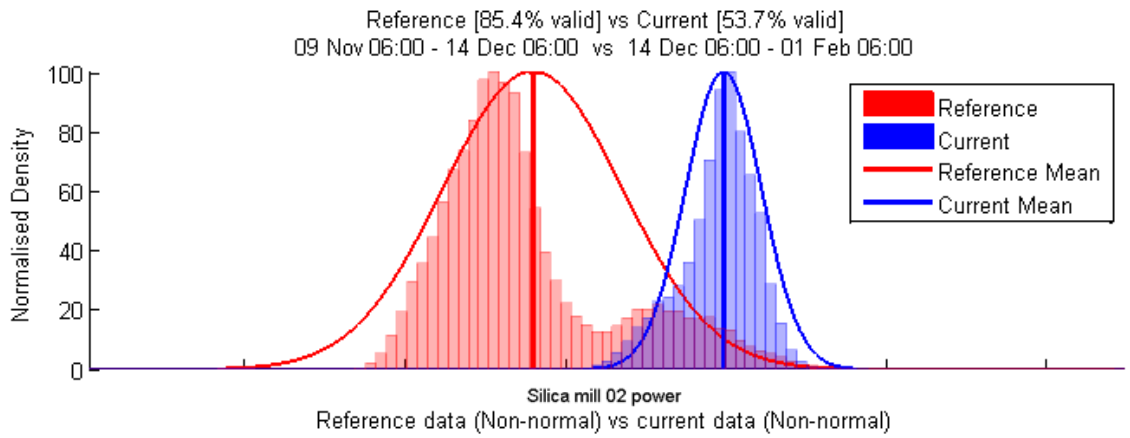

Figure 114: **OPM report: distribution comparison of silica mill 02 power (reference = class 2; current = class 3)**

For the evaluation of the silica mill performance, equipment stops and starts were also investigated. Equipment stops and starts are important to monitor as they can cause process instability and increase mechanical wear on the equipment. The number of stops is simply calculated by looking at the equipment run signal and applying a threshold to limit the number of false stops/starts counted. False stops/starts occur when problems are experienced during the restart of equipment, with equipment tripping and thus showing false stops/starts.

For silica mill 01 (Table 15 & Figure 115) it can be seen that from class 1 to class 2 there was a decrease in both the availability of the mill and the number of stops, but an increase in the average stop stability index (higher being better), with the exact opposite taking place when moving from class 2 to class 3. However, for silica mill 02 (Table 16 & Figure 116) it can be seen that from class 1 to class 2 the availability of the mill stayed constant, but the number of stops increased significantly, resulting in a decrease in the average stop stability index, with a comparative drop in both the availability and total number of stops when moving from class 2 to class 3. This has a significant effect on the performance of the mill, destabilising the process, with each mill stop causing up to eight hours of process instability.

Table 15: **OPM report: silica mill 01 equipment stops statistics**

	Class 1	Class 2	Class 3
Availability %	86.96	65.11	69.86
Total number of stops	25	12	31
Average stop stability index	0.52	0.70	0.34

Stellenbosch University http://scholar.sun.ac.za

Table 16: **OPM report: silica mill 02 equipment stops statistics**

	Class 1	Class 2	Class 3
Availability %	86.06	86.61	56.28
Total number of stops	23	52	32
Average stop stability index	0.55	0.21	0.23

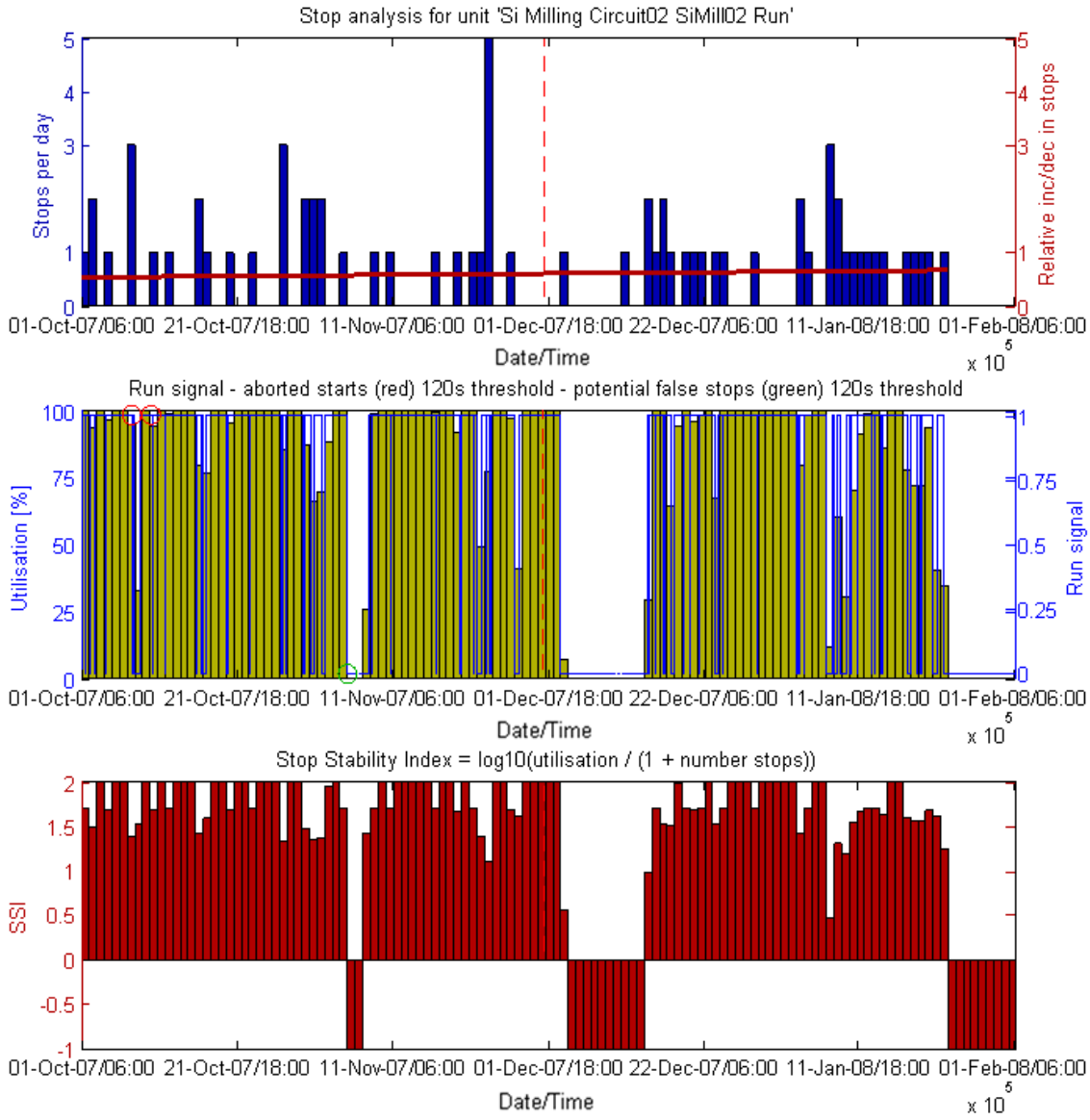

Figure 115: **OPM report: silica mill 01 equipment stops analysis**

Stellenbosch University http://scholar.sun.ac.za

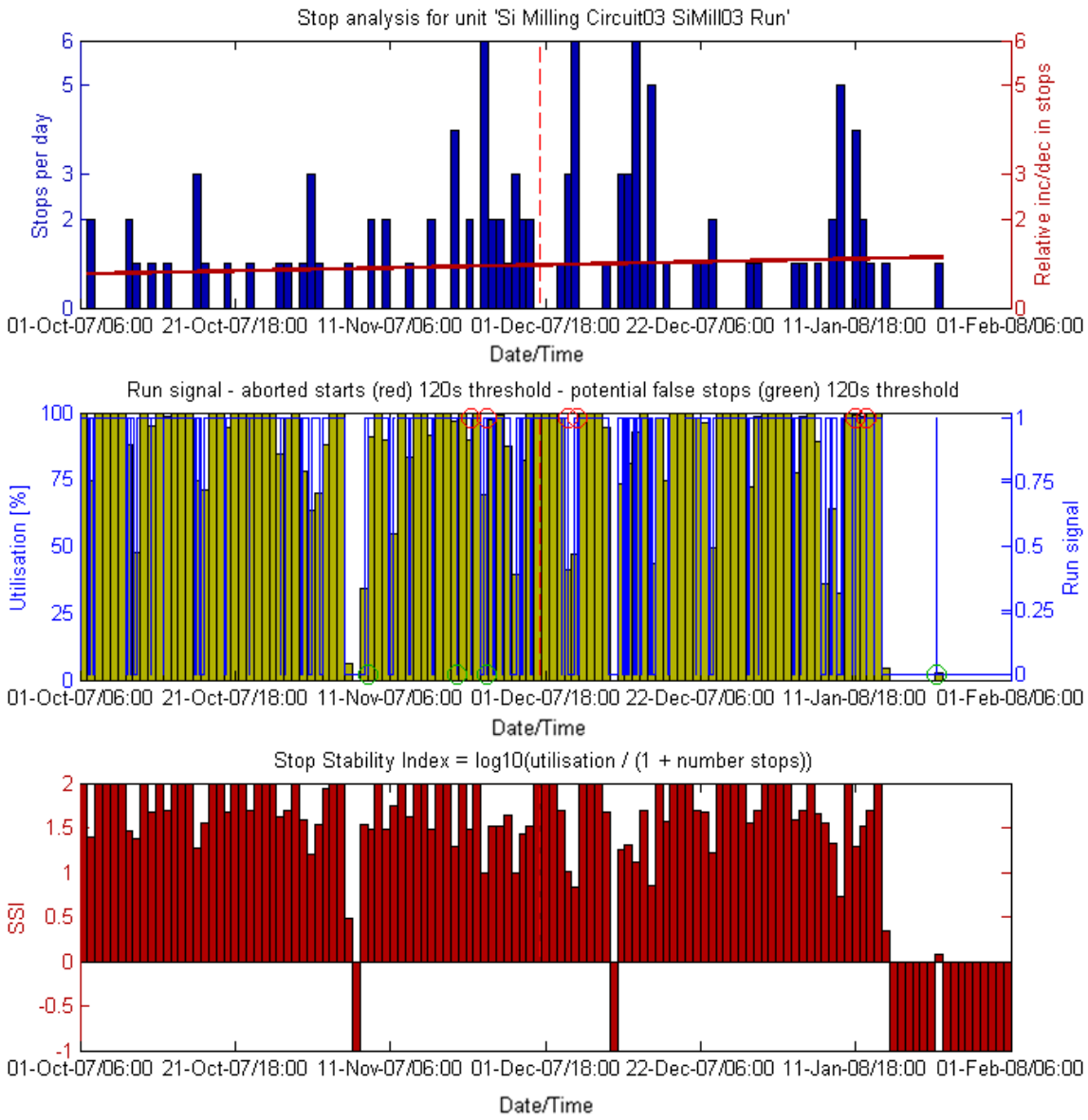

Figure 116: **OPM report: silica mill 02 equipment stops analysis**

With no significant shift in the primary flotation (rougher) tail grind and a significant increase in the power drawn by both silica mills which coincided with a coarsening of the silica circuit tail grind, fundamentally it can be concluded that there must have been an increase in the feed rate (solids feed rate) to the silica circuit: probably overloading the silica circuit, hence the higher silica mill power drawn and the coarsening of the silica circuit tail grind. From a process causality map (Figure 64) perspective, this could only be as a result of a shift in the chrome classification performance.

Stellenbosch University http://scholar.sun.ac.za

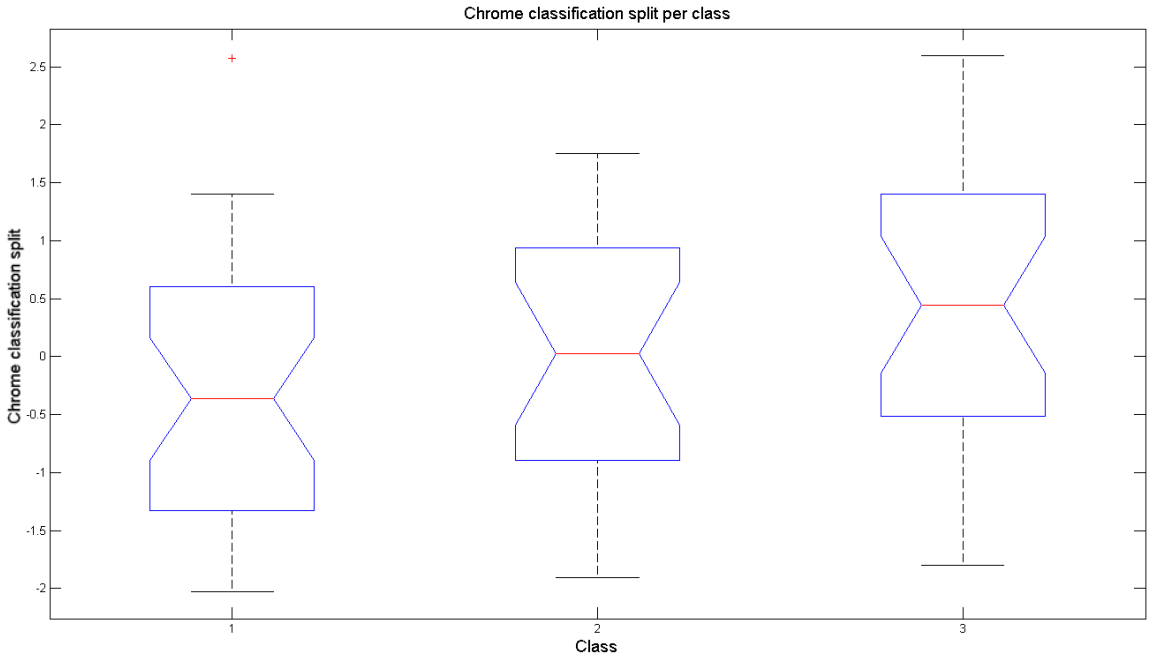

Figure 117: **Median significance: chrome classification mass split**

Dividing the chrome classification mass split data into classes similar to the recovery data and subjecting these classes of data to the median significance (Figure 117) and one-way ANOVA analysis, it was found that there was a significant difference between the classes at a significance level of less than 0.05. This would seem to confirm that a shift has occurred in the chrome classification performance which would have resulted in a shift in recovery.

Although a very good correlation has been found between the change points detected in the recovery variable and the chrome classification mass split data classes, it would again be interesting to see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others.

For the chrome classification mass split variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), significant change points were detected by the Bayesian probability change point detection algorithm (Figure 118) at sample 55 (15 day subsequent to the first recovery change point), and at sample 80 (5 days subsequent to the second recovery change point). An additional, seemingly significant change point was also detected by both the SSA and ELM SSA change point detection algorithms at sample 85. Although none of these change points coincide exactly with the recovery data change points, together with the median significance results (Figure 117) it can be concluded that the chrome classification mass split data can be classed

Stellenbosch University http://scholar.sun.ac.za

similarly to the recovery data. However, should the root cause of the decrease in recovery prove difficult to determine, using this classification of the chrome classification mass split data, the data can be re-classed prior to subsequent analyses.

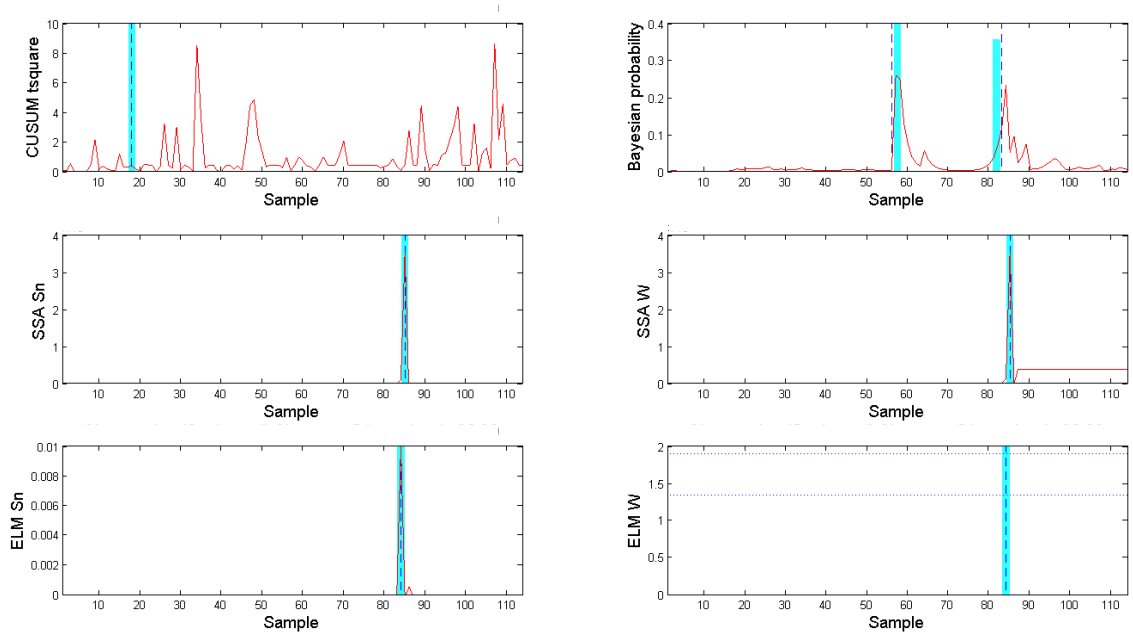

Figure 118: **Change point detection: chrome classification mass split at a confidence level of 0.99**

At this stage of the analysis it is evident that a shift in the chrome classification performance (in part an increase in the chrome classification mass split) and an increase in the feed rate (solids feed rate) to the silica circuit contributed to a shift in the silica circuit performance, primarily resulting in an increase in the silica circuit tails grade and a coarser silica circuit tails grind, which in turn contributed to a shift in the final tails grade and grind which contributed to a decrease in the recovery. Consequently, the focus is next on identifying the drivers that have caused a shift in the chrome classification performance, ultimately resulting in a decrease in the recovery.

7.1.7 Drivers: Chrome circuit grade and grind

Although the shift in the chrome circuit performance was much less prominent than the shift in the silica circuit performance, and therefore less important when considering the decrease in the recovery, a brief investigation into the chrome circuit performance is done (Figure 119). The investigation is, however, solely based on standard OPM analyses, simply reviewing some of the comparison graphs to support previous findings.

Stellenbosch University http://scholar.sun.ac.za

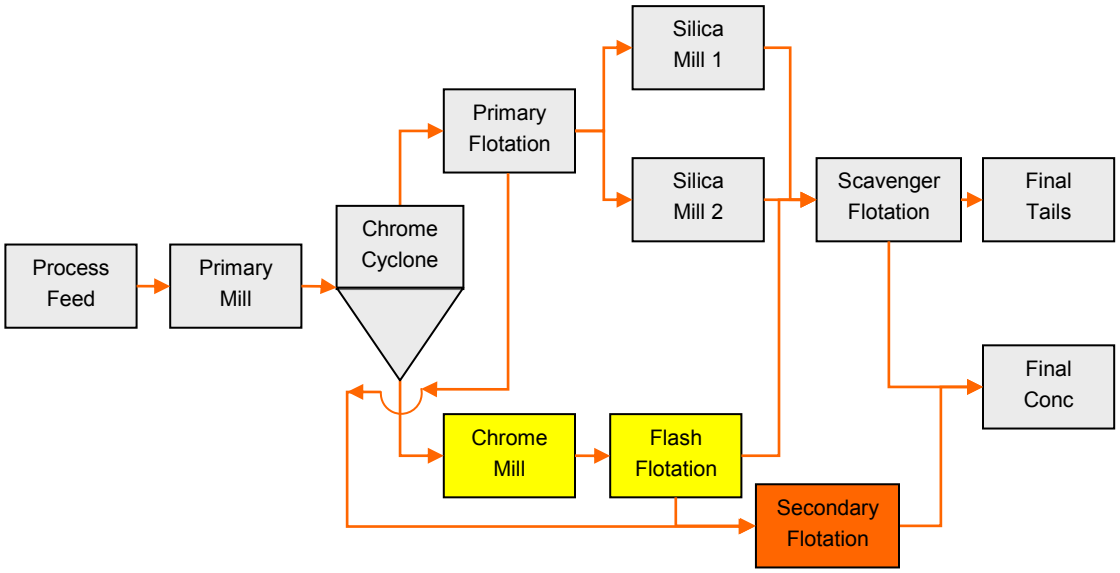

Figure 119: **Drivers: chrome circuit**

Firstly, the chrome flotation performance is reviewed by looking at the flash flotation circuit. From a process causality map (Figure 63) perspective, flotation performance is a function of mass pull and feed. Unfortunately, no reliable feed measurements are available for the flash float circuit and only the mass pull component can be looked at. Mass pull in turn is a function of froth depth, approximated by the pulp level measurement, and air flow rate.

From the flash float level KPIs analysis (Figure 120), it can be seen that for the same level process measurement for all 3 classes, the valve output (MV) has decreased significantly for class 2 and class 3 compared to class 1.

Stellenbosch University http://scholar.sun.ac.za

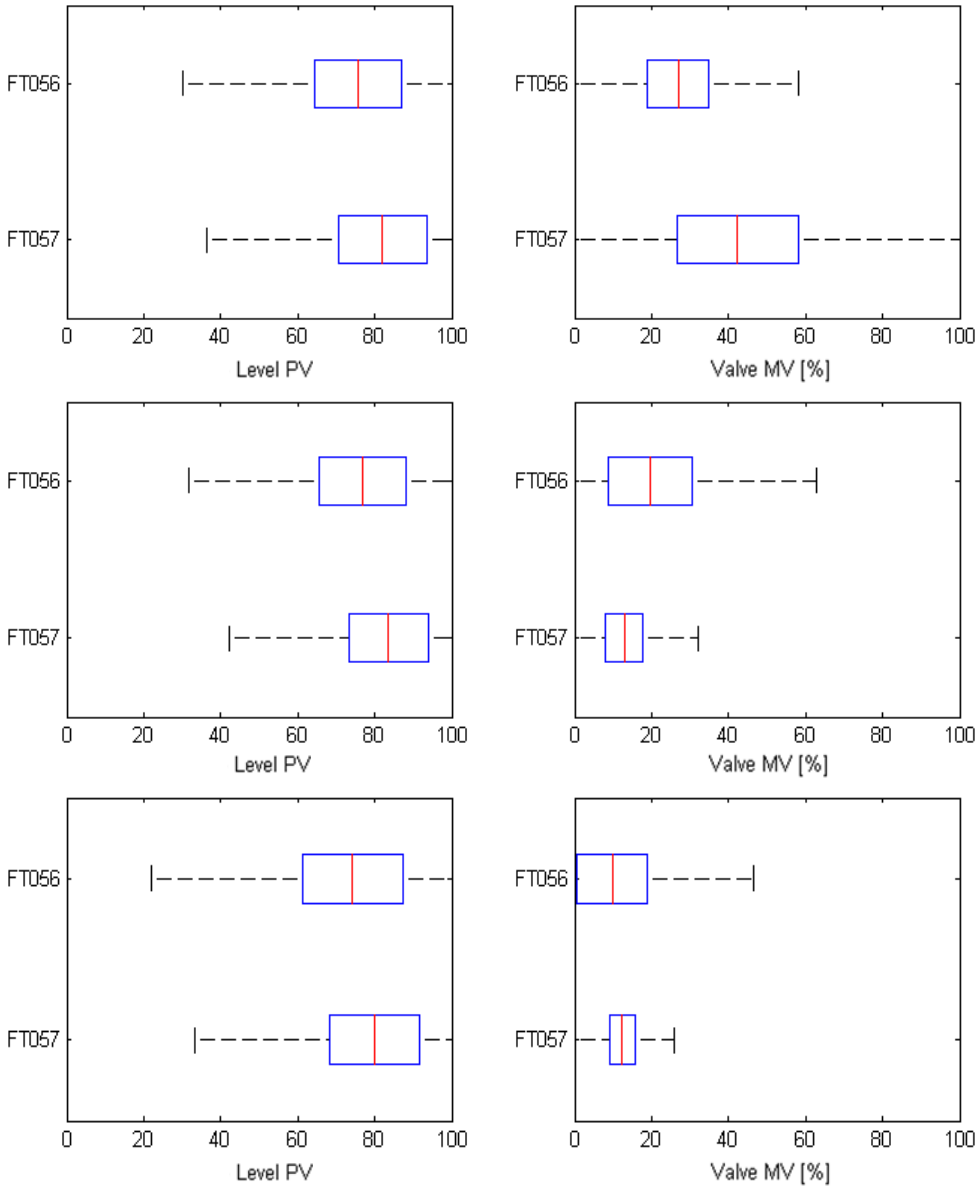

Figure 120: **OPM report: class 1, 2 and 3 (top to bottom) flash float level KPIs**

Similarly, from the flash float air KPIs analysis (Figure 121), it can be seen that for the same air flow rate process measurement for all 3 classes, the valve output (MV) has decreased significantly for class 2 and class 3 compared to class 1. For the valve air flow rate to stay the same at a lower valve output, the material in the flotation cell had to become less dense (having a larger water component), which could be attributed to a shift in the chrome classification performance.

Stellenbosch University http://scholar.sun.ac.za

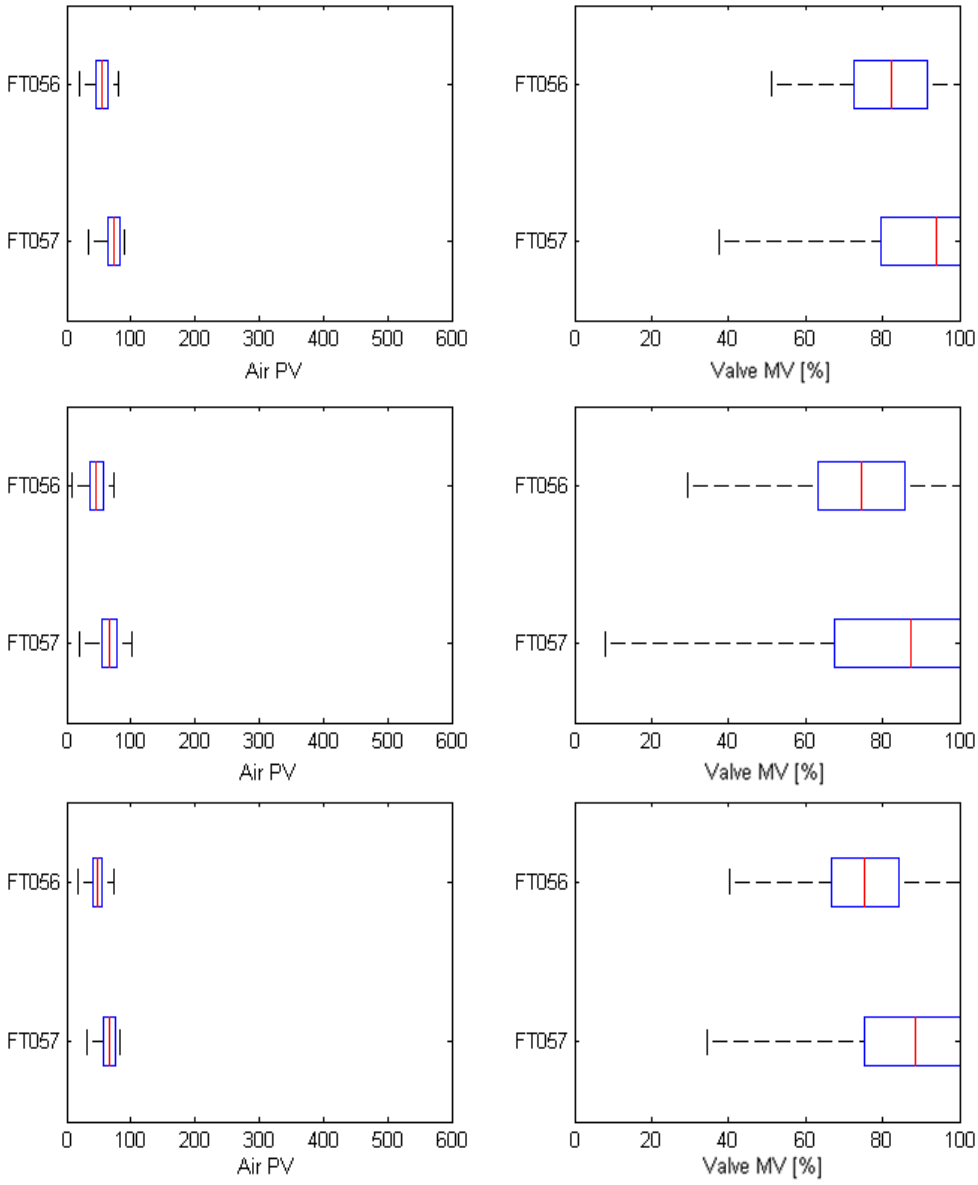

Figure 121: **OPM report: class 1, 2 and 3 (top to bottom) flash float air KPIs**

Secondly, the chrome milling performance is reviewed by looking at the chrome mill. From a process causality map (Figure 60) perspective, the chrome milling is a function of variables such as mill power, mill load, mill throughput and mill feed PSD. Unfortunately, most of these measurements are not reliably available and only mill power and mill load will be analysed. For this purpose, a 2-D histogram plot is used. 2-D histogram plots are a graphical means of relating two variables, with more dense areas being displayed as red. Scatter plots, with time indicated using coloured markers, are plotted in support of the 2-D histogram plots. Whereas the 2-D histogram plots show a graphical non-statistical clustering of the data, the scatter plots show the movement of data over time as indicated by coloured markers (red

Stellenbosch University http://scholar.sun.ac.za

markers representing the oldest data and blue markers representing the most recent data). High and low control limits are also indicated on these plots (green lines indicating the low control limits and purple lines indicating the high control limits).

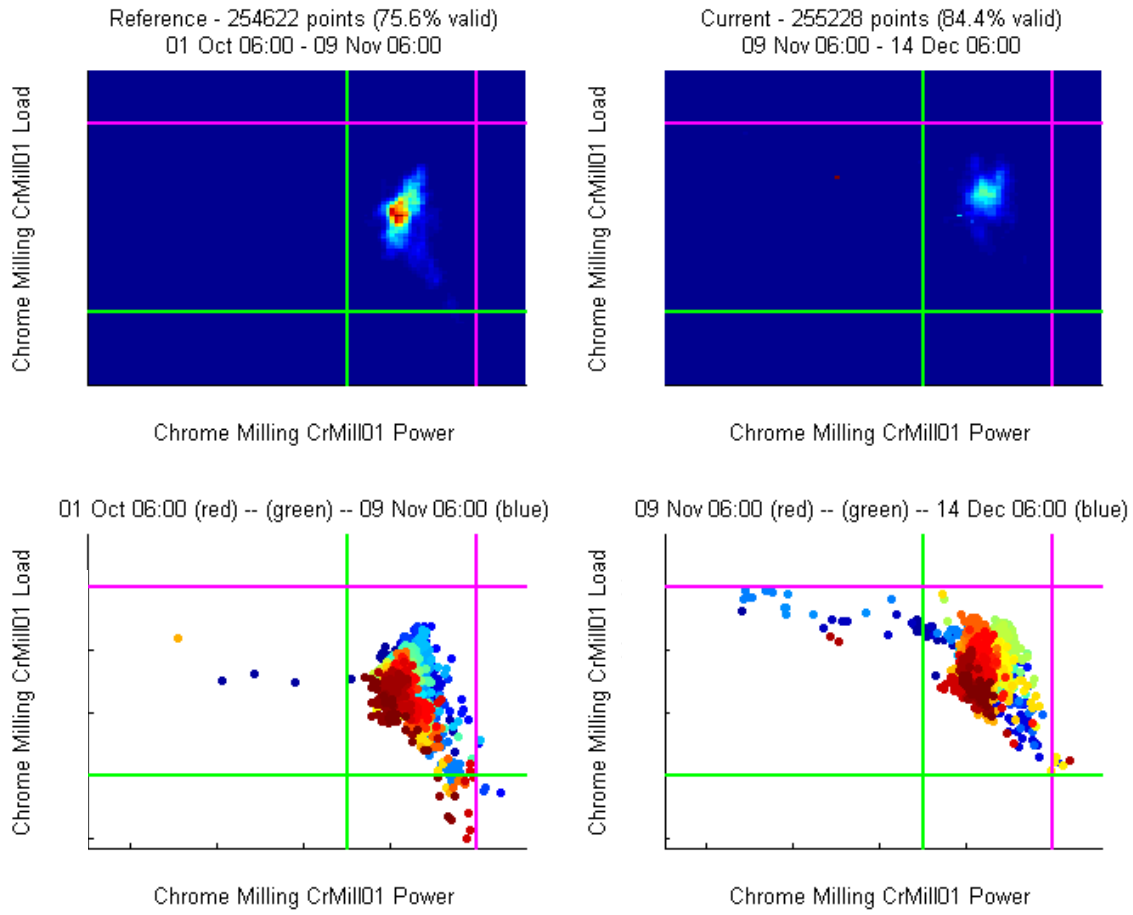

Figure 122: **OPM report: 2-D histogram plot of power versus load (reference = class 1; current = class 2)**

From the chrome mill power versus load 2-D histogram plot (Figure 122) for class 1 versus class 2, it is evident that there was an increase in both the chrome mill power and load. Fundamentally this would be due to an increased feed rate, resulting from an increased mass split to the chrome circuit, to the chrome mill. From this, together with the flash float analysis findings, it may then be assumed that the mass split to the chrome circuit is also higher due to a higher water split to the chrome circuit (resulting from a decrease in the chrome classification performance/efficiency). In turn, this would cause the increase in solids feed rate to the silica circuit that was noted.

Stellenbosch University http://scholar.sun.ac.za

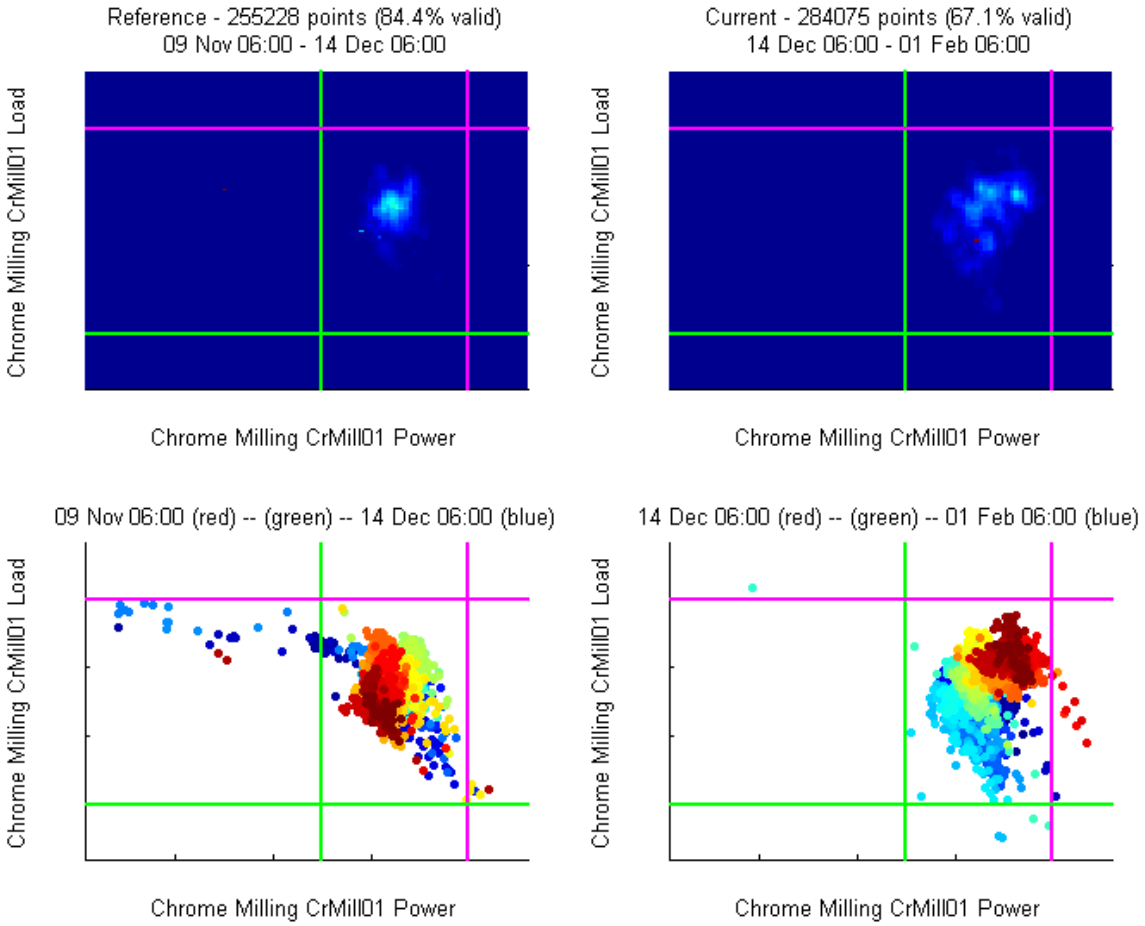

Figure 123: **OPM report: 2-D histogram plot of power versus load (reference = class 2; current = class 3)**

From the chrome mill power versus load 2-D histogram plot (Figure 123) for class 2 versus class 3, a subsequent decrease in both the chrome mill power and load can be seen. This is indicative of the chrome circuit returning to its initial state. With both the chrome circuit grade and grind being largely unaffected by these changes in the chrome circuit it is safe to assume that there was sufficient excess capacity in the chrome circuit to absorb the increased mass split to the chrome circuit.

7.1.8 Drivers: Chrome classification

Following the findings of the silica circuit and chrome circuit analysis, the focus shifts to finding the drivers that have caused a shift in the chrome classification performance (Figure 124). From a process causality map (Figure 61) perspective, the chrome classification (cyclone) performance is a function of the operating pressure and the feed PSD. Unfortunately, no reliable feed PSD measurements are available

Stellenbosch University http://scholar.sun.ac.za

for the cyclone and only the pressure component, being a function of feed flow rate and feed density, will be looked at.

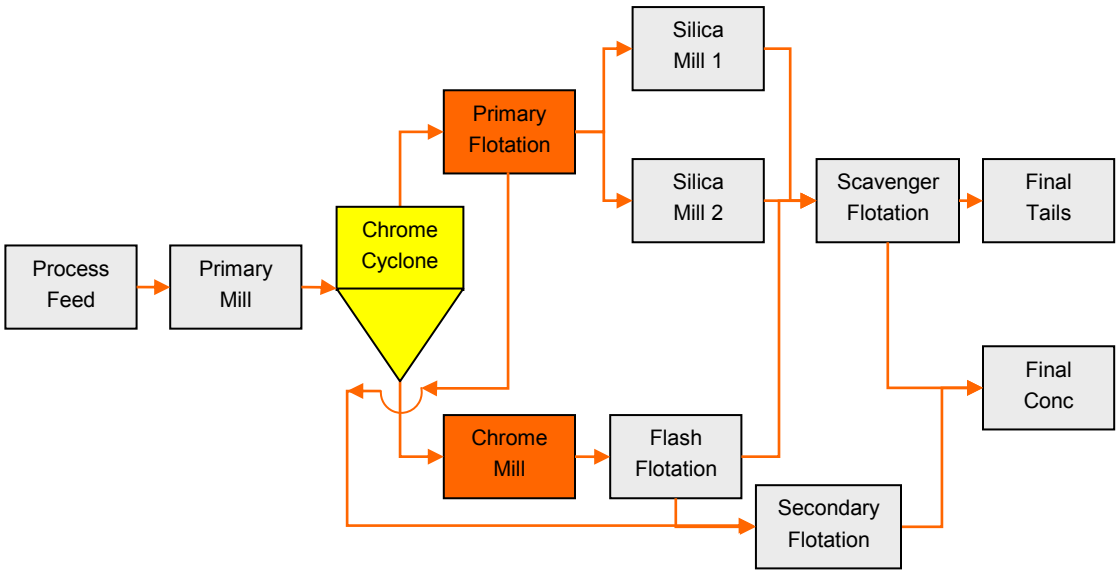

Figure 124: **Drivers: chrome classification**

As before, fault detection models were constructed using only the class 1 reference data from the chrome classification drivers data set with performance metrics calculated for the class 2 changeover and class 3 fault data (Figure 125). However, it was found that the reliability of none of the performance metrics exceeded 65% (Figure 125). This could be a direct consequence of any of the following: (1) some of the other process variables found in the combined data set are probably better indicators of the fault condition in the process compared to the process variables in the chrome classification drivers data set, (2) the magnitude of the fault condition as represented by the selected process variables is too small to result in a measurable process performance degradation over the evaluation period (variation still within variation associated with common cause), (3) the recovery data classes do not align precisely with the chrome classification data classes, or (4) other process variables or indicators exist that are not being measured that could have a positive effect on the reliability metric.

Stellenbosch University http://scholar.sun.ac.za

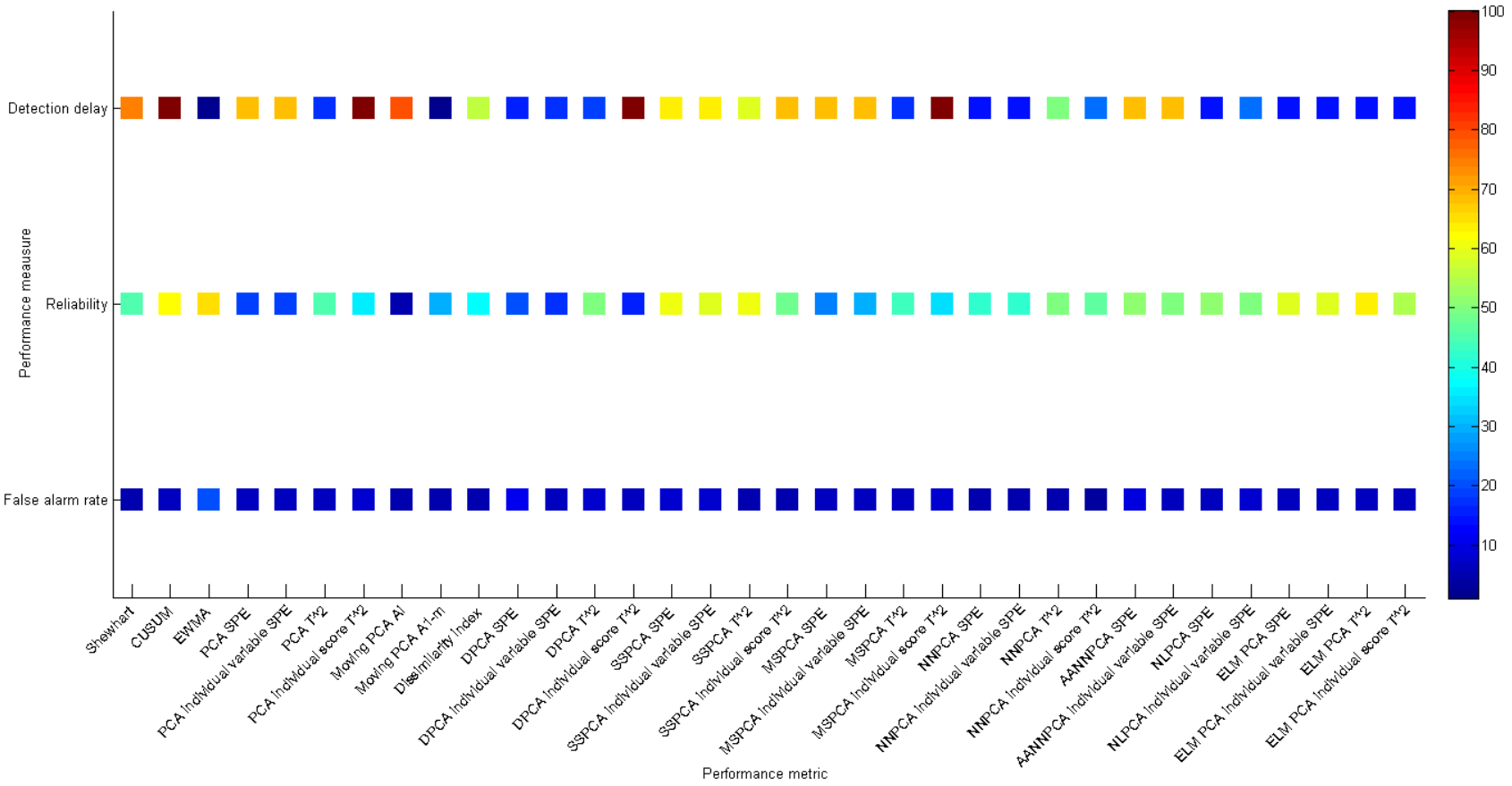

Figure 125: **Statistical data-based fault detection: chrome classification – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 126) is made to determine if the chrome classification performance drivers can be used to distinguish between the different recovery classes.

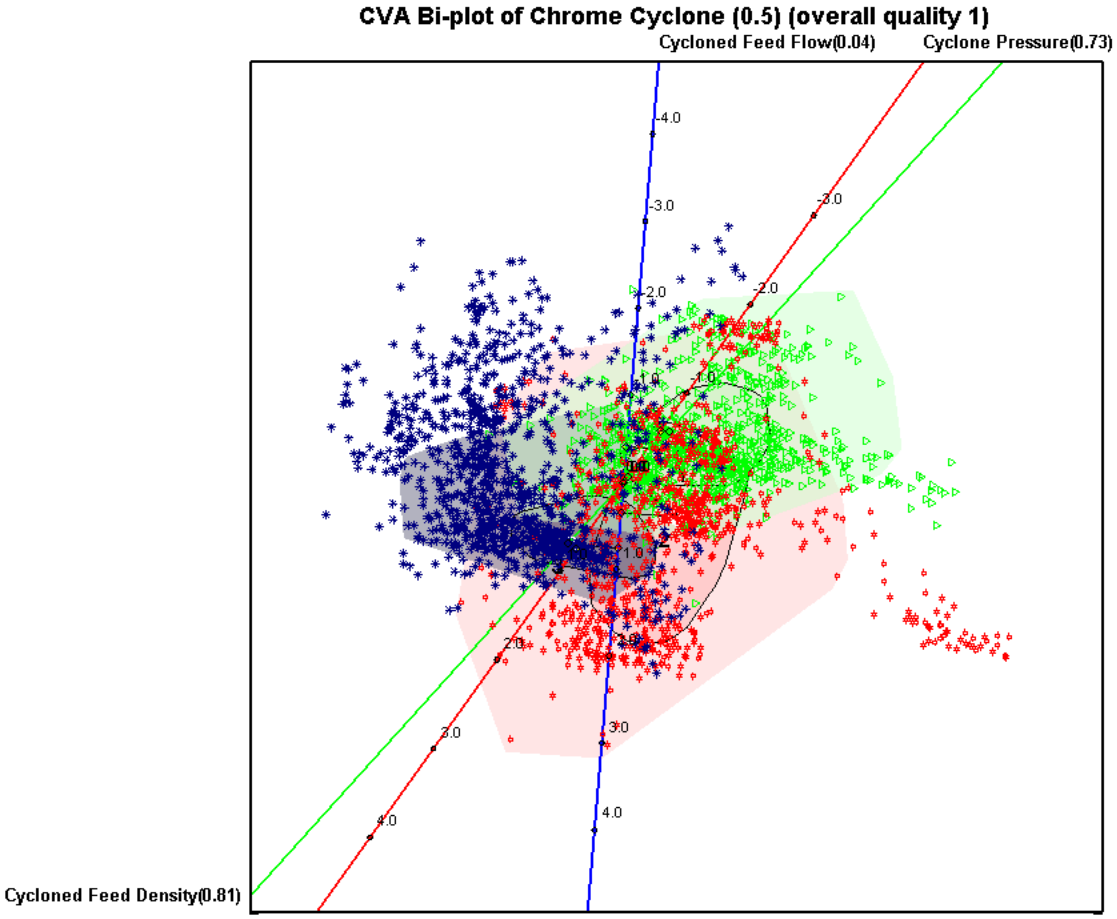

Figure 126: **CVA biplot: chrome classification (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

From the CVA biplot it is evident that the classes are not well separated, especially class 1 and class 2, with the alpha bags overlapping each other even at an alpha value of 0.5. With the chrome classification cyclone feed flow rate variable axis have a very low predictivity, significant and reliable conclusions can only be drawn with regards to the discriminatory ability of the chrome classification cyclone pressure and chrome classification cyclone feed density.

It can further be seen from the CVA biplot that there is a very high degree of correlation between the variables related to the chrome classification cyclone performance, as is evident from the small angle between the biplot axes for these measurements. This supports the structure of the fundamentally derived cyclone process causality map. Visually, the CVA biplot shows that there is no single good

Stellenbosch University http://scholar.sun.ac.za

differentiator between class 2 and class 3 of the data. However, when moving from class 1 to class 2 or class 3, the chrome classification cyclone pressure and the chrome classification cyclone feed density are seen to be good differentiators between the classes. From this visual inspection and analysis of the CVA biplot results it would seem as if, although class 2 and class 3 of the data are not similar, they are similarly different from class 1 in respect of the chrome classification cyclone pressure and the chrome classification cyclone feed density.

As before, the CVA biplot analysis is supported by the variable importance analysis. From the variable importance analysis on average over all 3 classes (Figure 127) the main differentiators between the recovery classes are the chrome classification cyclone pressure and chrome classification cyclone feed flow rate (this being contrary to the CVA biplot analysis where the chrome classification cyclone feed flow rate had a very low axis predictivity). Moving only from class 1 to class 2 (Figure 128) or class 2 to class 3 (Figure 129), the chrome classification cyclone feed flow rate is the most significant driver followed by the chrome classification cyclone pressure. These results correspond fairly well with the findings from analysing the CVA biplot, with the chrome classification cyclone feed flow rate being better represented in the variable importance analysis compared to the CVA biplot analysis.

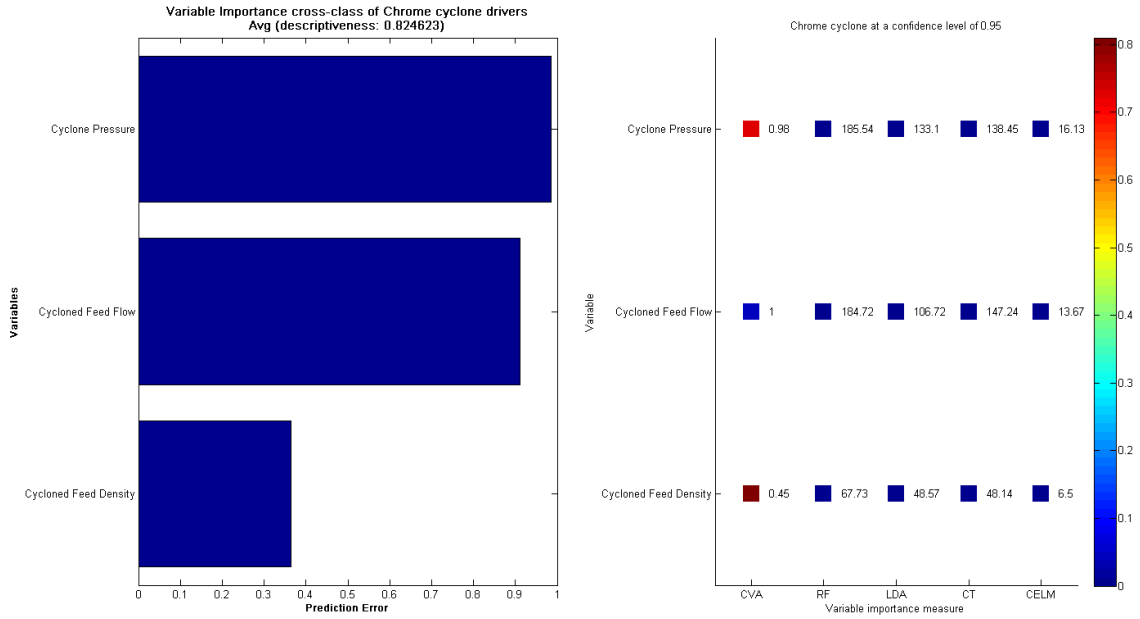

Figure 127: **Variable importance class 1-2-3: chrome classification at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

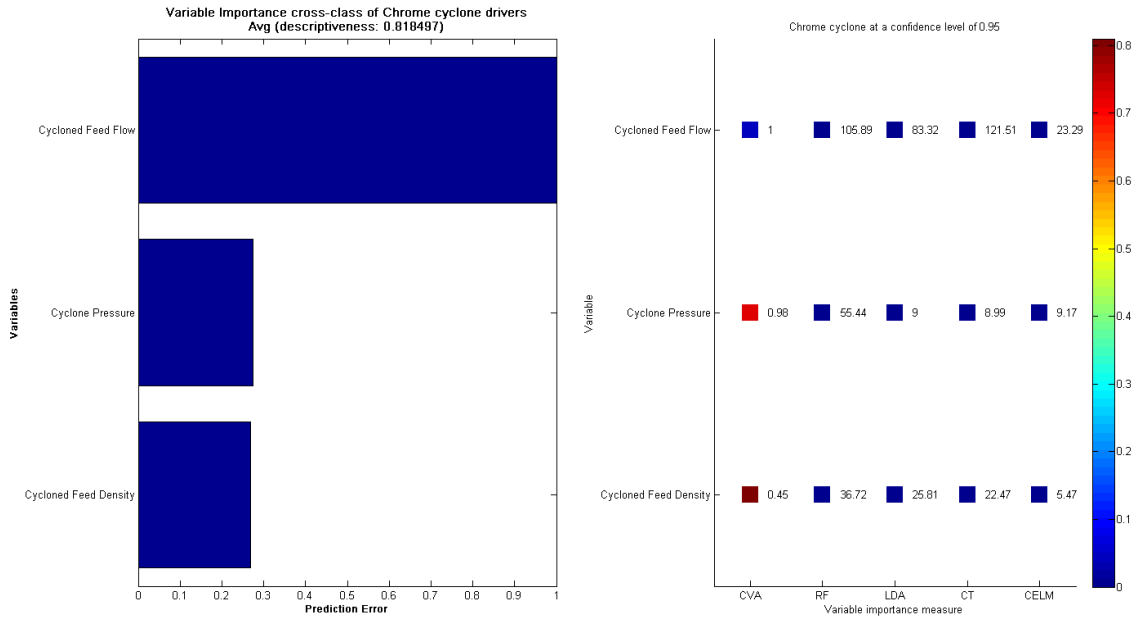

Figure 128: **Variable importance class 1-2: chrome classification at a confidence level of 0.95**

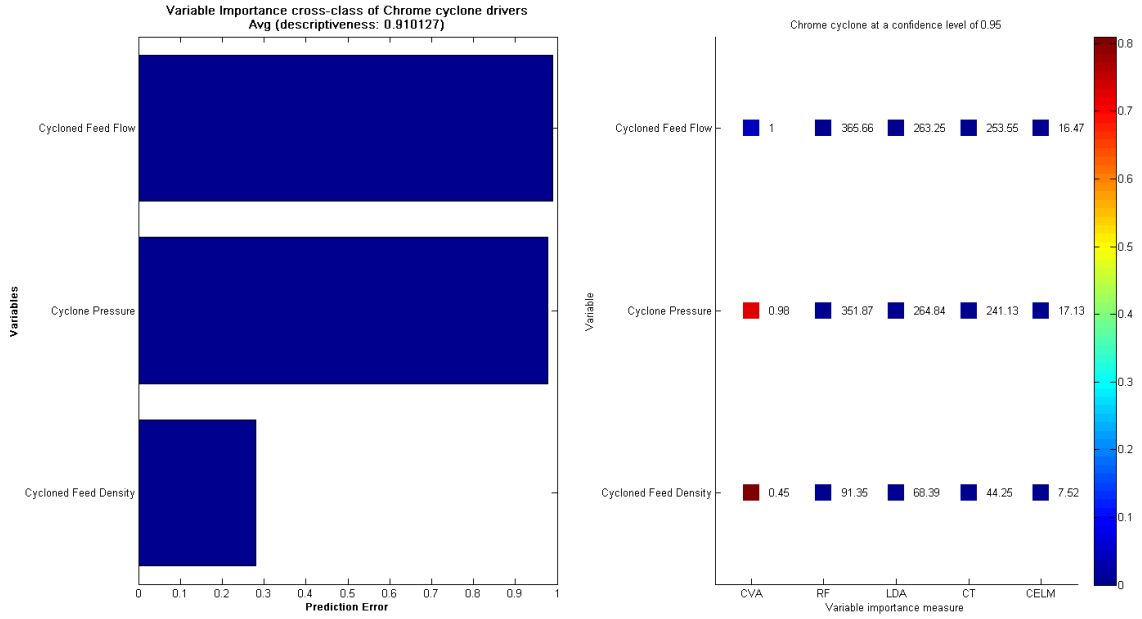

Figure 129: **Variable importance class 2-3: chrome classification at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

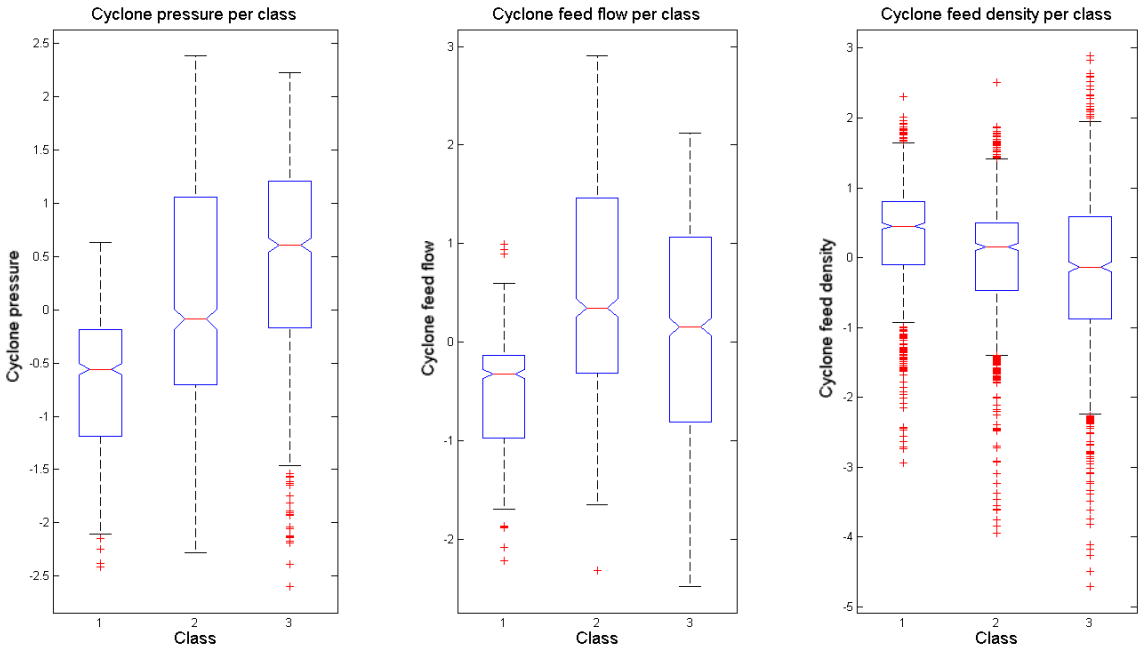

Figure 130: **Median significance: cyclone pressure, cyclone feed flow rate and cyclone feed density**

Subjecting the 3 cyclone pressure, cyclone feed flow rate and cyclone density classes of data to the median significance (Figure 130) and one-way ANOVA analysis, it was found that there was a significant difference between the classes for all these variables at a significance level of less than 0.05. It can thus be concluded that there is a good link between the chrome classification mass split classes and the chrome classification cyclone operating variables, especially the cyclone pressure, consistently increasing from one class to the next, and the cyclone density, consistently decreasing from one class to the next.

Although a good correlation has been found thus far between the change points detected in the recovery variable and potential important variables related to the chrome classification performance, it would be interesting to see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others. Whereas the initial data was sampled shiftly (8 hourly), the chrome classification data was sampled 2 minutely. Subsequently, the comparative recovery change points for this data set can be found at sample 891 (08/11/2007) and 1731 (13/12/2007).

For the chrome classification cyclone pressure variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), none of the significant change points detected in the recovery variable are noticeable in the change point detection results

Stellenbosch University http://scholar.sun.ac.za

(Figure 131). This can in part explain the significant overlap that was present in the CVA biplot (Figure 126) between class 1 and class 2, especially considering the change point that was identified by the nearest-neighbours CUSUM and ELM SSA techniques at around sample 1350 (midway through the class 2 data set). Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

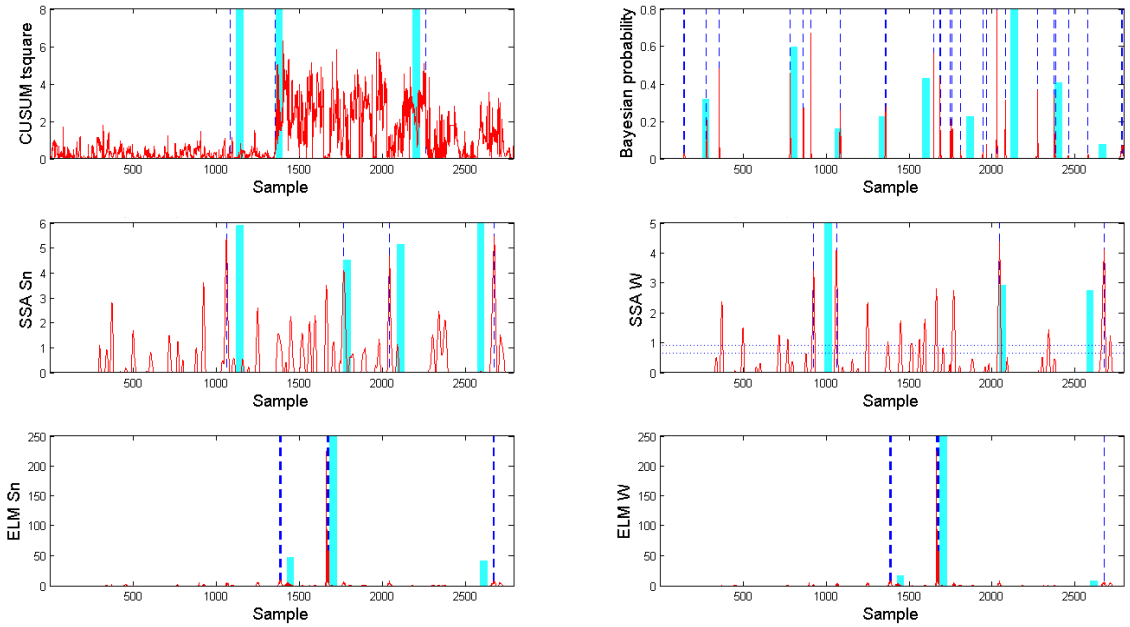

Figure 131: **Change point detection: cyclone pressure at a confidence level of 0.99**

For the chrome classification cyclone feed flow rate variable (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), none of the significant change points detected in the recovery variable were explicitly detected (Figure 132). As with the chrome classification cyclone pressure variable, this can in part explain the significant overlap that was present in the CVA biplot (Figure 126) between class 1 and class 2, especially considering the change point that is visually identifiable in the nearest-neighbours CUSUM and ELM SSA technique results at around sample 1350 (midway through the class 2 data set). Furthermore, although again not explicitly detected, the decrease in the nearest-neighbours CUSUM T^2 statistic coincides with the recovery change point between class 2 and class 3, supporting the evidence thus far that the shift in the chrome classification cyclone feed flow rate variable contributed to the shift in the recovery variable. Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

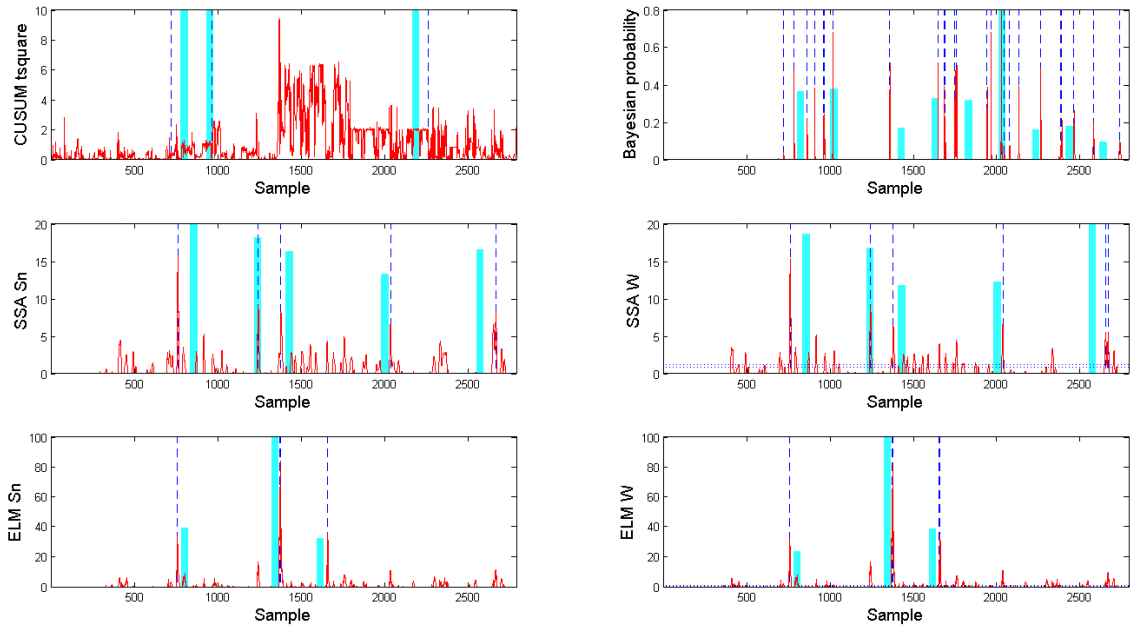

Figure 132: **Change point detection: cyclone feed flow rate at a confidence level of 0.95**

For the chrome classification cyclone feed density variable (autocorrelated, not normally distributed, exponentially distributed, SSA suggested change point detection technique), none of the significant change points detected in the recovery variable were explicitly detected (Figure 133). The Bayesian probability change point detection algorithm did, however, detect numerous significant change points. This phenomenon can be ascribed to the fact that the density measurement is a very unstable measurement, regularly shifting due to process disturbances, compared to much better controlled measurements, such as flow rates. Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

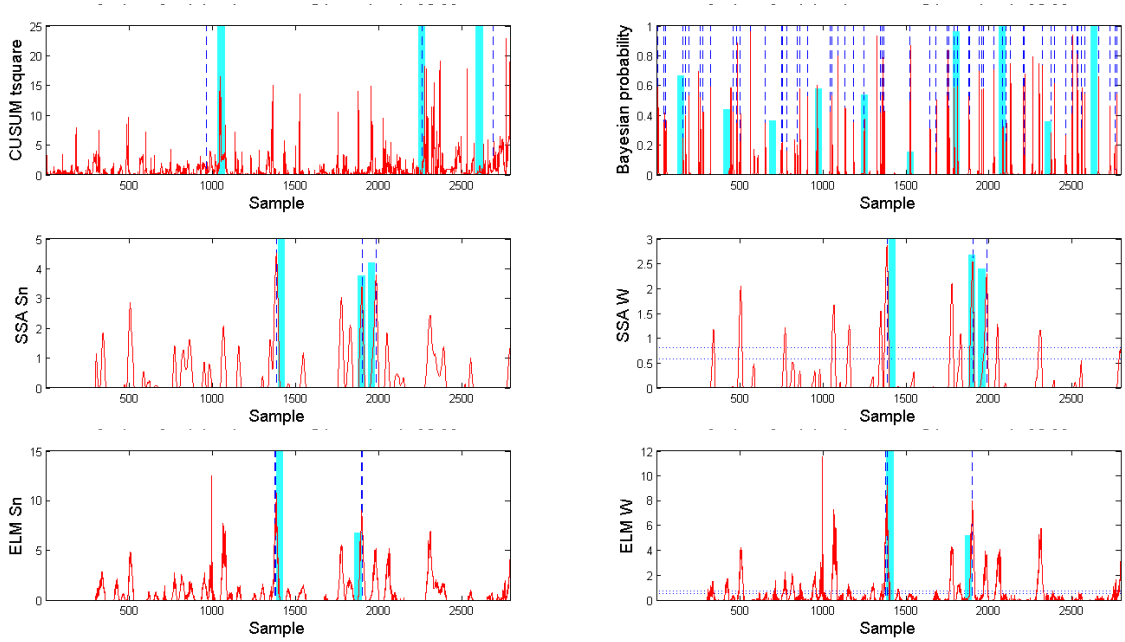

Figure 133: **Change point detection: cyclone feed density at a confidence level of 0.95**

The significance of the changes in the variables related to the chrome classification cyclone, especially the chrome classification cyclone pressure, can be better visualised using higher frequency data. For this purpose OPM graphs, in the form of distribution plots, are used, not only showing the operating position of the variable but also giving an indication as to the stability of the variable.

For the chrome classification cyclone pressure distribution plot it can be seen that only the class 1 data has a unimodal distribution (Figure 134), with both the class 2 and class 3 data having bimodal distributions with significant overlap between the classes (Figure 135). This further confirms the CVA biplot findings that there is significant overlap between the chrome classification cyclone pressure data classes and that these classes do not align precisely with the recovery data classes. This could be indicative of the fact that although the shift in chrome classification performance contributed to the shift in the recovery, it might not be the only factor responsible.

Stellenbosch University http://scholar.sun.ac.za

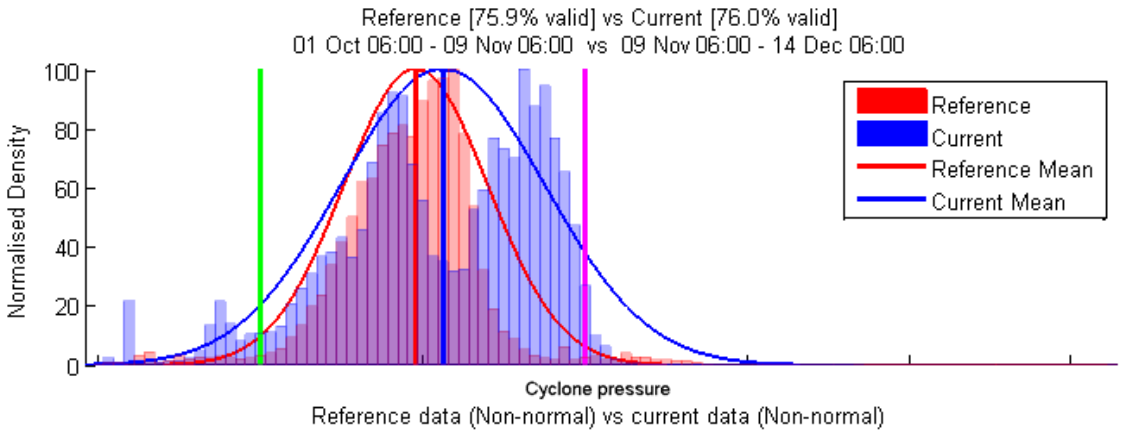

Figure 134: **OPM report: distribution comparison of cyclone pressure (reference = class 1; current = class 2) also indicating the upper (magenta) and lower (green) operating limits**

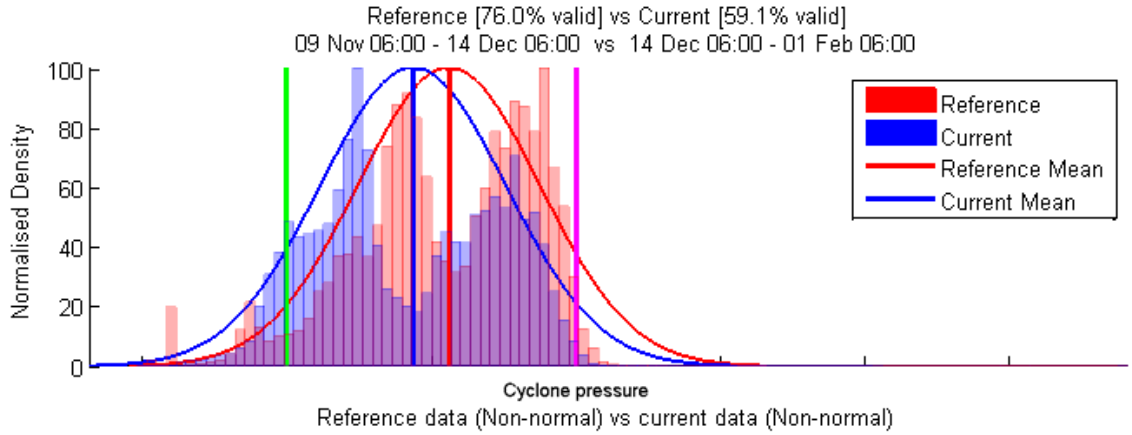

Figure 135: **OPM report: distribution comparison of cyclone pressure (reference = class 2; current = class 3) also indicating the upper (magenta) and lower (green) operating limits**

For the chrome classification cyclone feed flow rate distribution plot it can be seen that again the class 1 data has a unimodal distribution (Figure 136), with the class 2 data having a bimodal distribution and both the class 2 and class 3 data having negative skewness with some overlap between the classes (Figure 137). This shows that the changes in the chrome classification feed flow rate was also not solely responsible for the changes in the chrome classification pressure but that changes in the chrome classification feed density (as per the cyclone process causality map) also played a role.

Stellenbosch University http://scholar.sun.ac.za

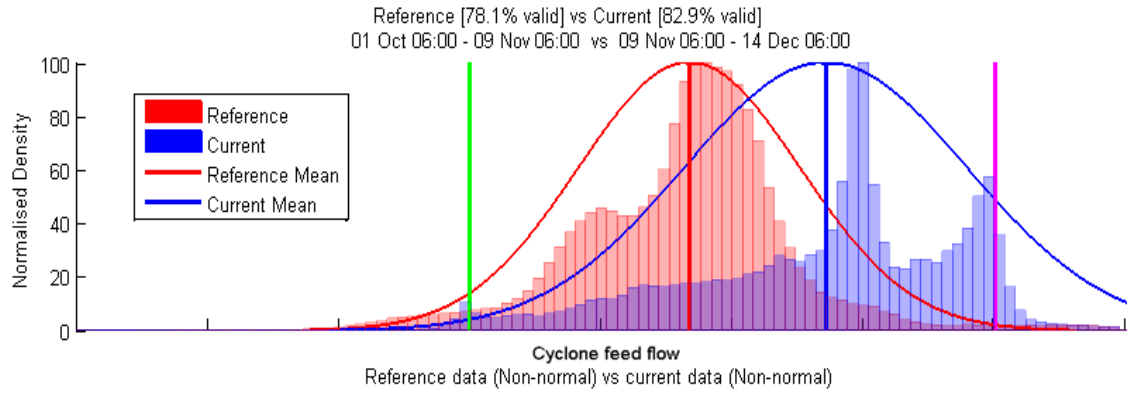

Figure 136: **OPM report: distribution comparison of cyclone feed flow rate (reference = class 1; current = class 2) also indicating the upper (magenta) and lower (green) operating limits**

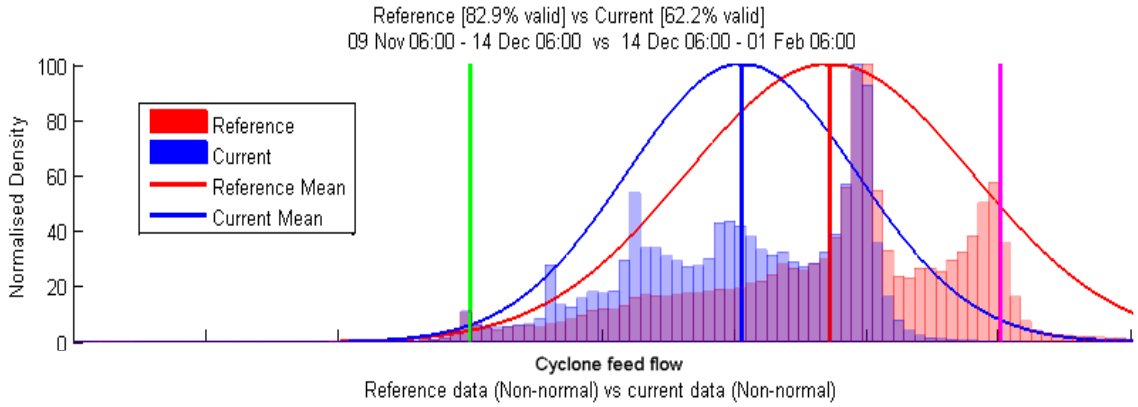

Figure 137: **OPM report: distribution comparison of cyclone feed flow rate (reference = class 2; current = class 3) also indicating the upper (magenta) and lower (green) operating limits**

For the chrome classification cyclone feed density distribution plot it can be seen that both the class 1 and class 2 data has a Gaussian distribution (Figure 138), with only the class 3 data having a bimodal distribution (Figure 139). Whereas for the class 1 and class 2 data there is visually very little difference between the distributions, for the class 3 data it can be seen that there is a definite widening of the distribution, possibly resulting in a more unstable operation.

Stellenbosch University http://scholar.sun.ac.za

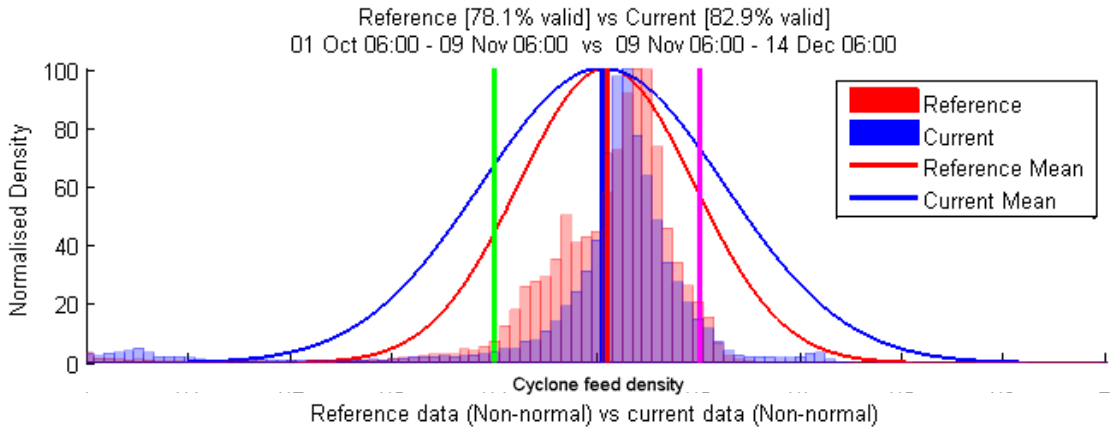

Figure 138: **OPM report: distribution comparison of cyclone feed density (reference = class 1; current = class 2) also indicating the upper (magenta) and lower (green) operating limits**

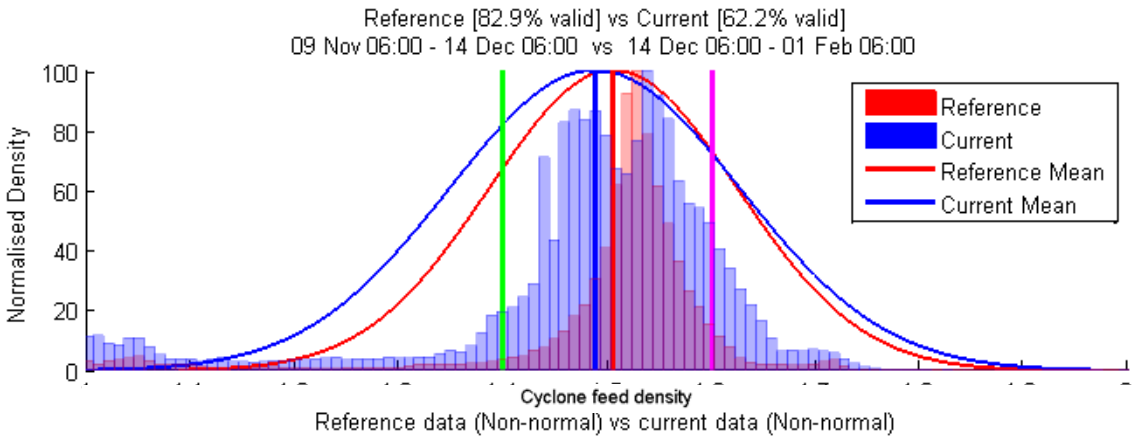

Figure 139: **OPM report: distribution comparison of cyclone feed density (reference = class 2; current = class 3) also indicating the upper (magenta) and lower (green) operating limits**

Similar to the 2-D histogram plots, parallel coordinate plots are generated to visualise relationships between multiple variables. Parallel coordinate plots allow you to visualise as many variables as you wish on a single graph, showing the values of different variables with a single view. This in turn allows for the isolation of problematic variables during visual fault detection and root cause analysis, given that the variables are arranged in the proper (causal) order. Simple line plots, with the progression of time indicated by the colour of the lines (red lines representing the oldest data and blue lines representing the most recent data), are plotted in support of the parallel coordinate plots. Control limits are also superimposed on these plots (thick yellow lines and markers).

Stellenbosch University http://scholar.sun.ac.za

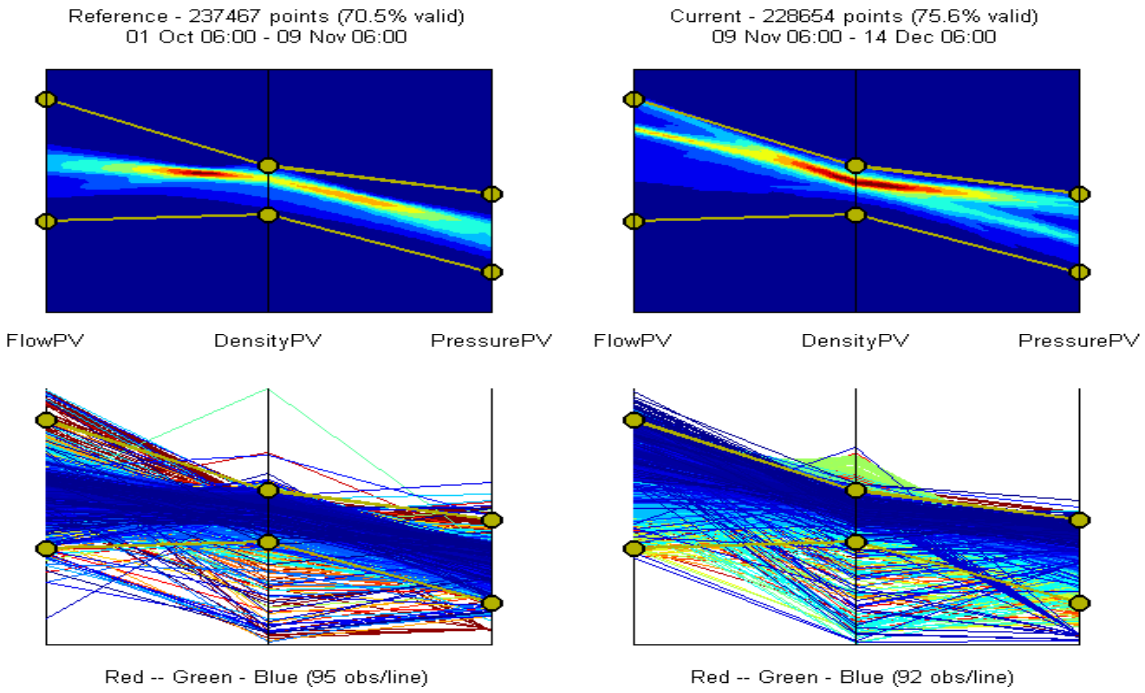

Figure 140: **OPM report: parallel coordinate plot of cyclone performance (reference = class 1; current = class 2)**

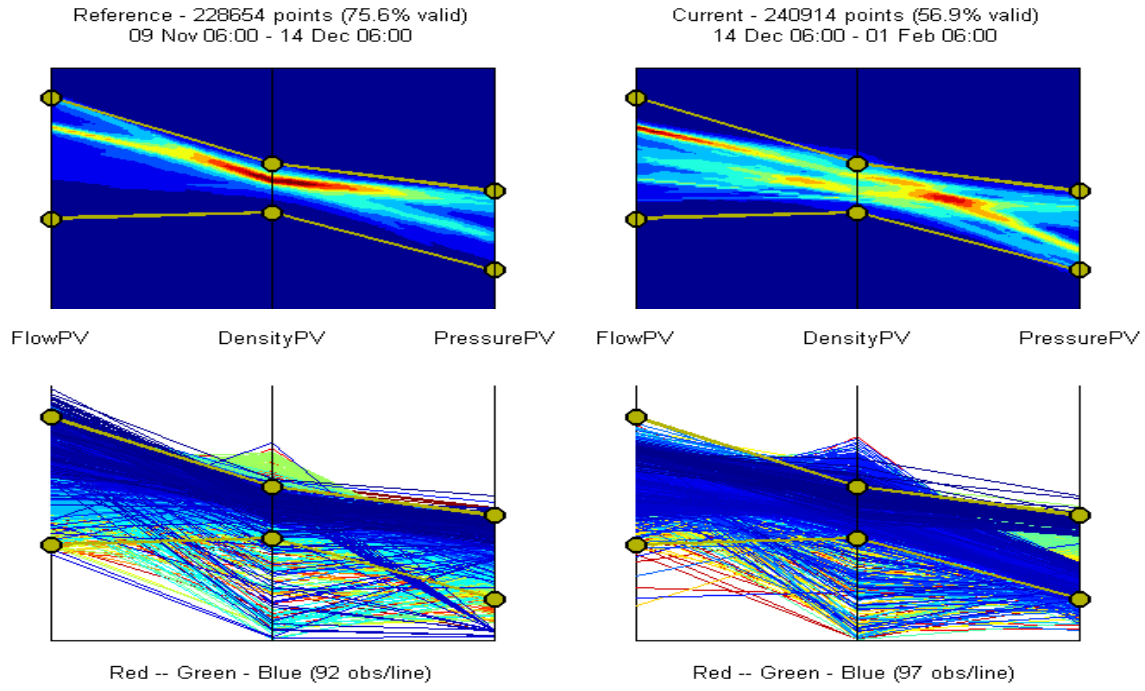

Figure 141: **OPM report: parallel coordinate plot of cyclone performance (reference = class 2; current = class 3)**

Stellenbosch University http://scholar.sun.ac.za

From the chrome classification cyclone parallel coordinate plots (Figure 140 and Figure 141) similar information is displayed as with distribution plots and, therefore, similar conclusions can be drawn, albeit for more variables on a single graph. Additional knowledge can, however, be gained from the parallel coordinate plots regarding the relationships between the variables:

- Average flows combined with high densities lead to average pressures (Class 1)
- High flows combined with high densities lead to high pressures (Class 2)
- Low flows combined with low densities lead to low pressures (Class 3)

This knowledge confirms the fundamental operation of the chrome classification cyclone and supports the understanding thereof.

At this stage of the analysis it is evident that the significant increase in the chrome classification cyclone pressure and an associated increase in the chrome classification cyclone feed flow rate (at least initially) and decrease in the chrome classification cyclone feed density, resulted in the increase in the chrome classification mass split, which in turn contributed to the decrease in the recovery. From a process causality map (Figure 64) perspective, it can therefore be concluded that a shift has occurred in the operation of the primary mill with the focus next being on identifying the drivers that have caused this shift in the primary mill, ultimately resulting in a decrease in the recovery.

7.1.9 Drivers: Primary mill

From a process causality map (Figure 60) perspective, the primary mill (Figure 142) is a function of variables such as mill power, mill load and in-mill density. Unfortunately, no reliable in-mill density measurement is available and consequently mill feed rate and mill inlet water ratio will be looked as being representative of in-mill density.

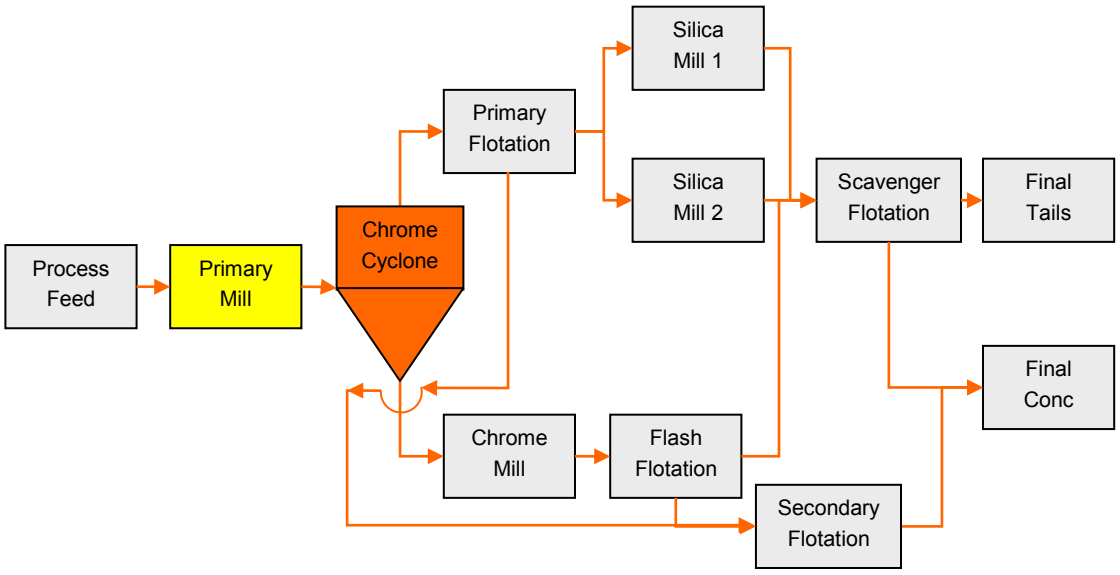

Figure 142: **Drivers: primary mill**

Stellenbosch University http://scholar.sun.ac.za

As before, fault detection models were constructed using only the class 1 reference data from the primary mill drivers data set with performance metrics calculated for the class 2 changeover and class 3 fault data. However, it was found that the reliability of none of the performance metrics exceeded 45% (Figure 143). As with the chrome classification drivers data set, this could be a direct consequence of any of the following: (1) some of the other process variables found in the combined data set are probably better indicators of the fault condition in the process compared to the process variables in the primary mill drivers data set, (2) the magnitude of the fault condition as represented by the selected process variables is too small to result in a measurable process performance degradation over the evaluation period (variation still within variation associated with common cause), or (3) the recovery data classes do not align precisely with the primary mill data classes.

Stellenbosch University http://scholar.sun.ac.za

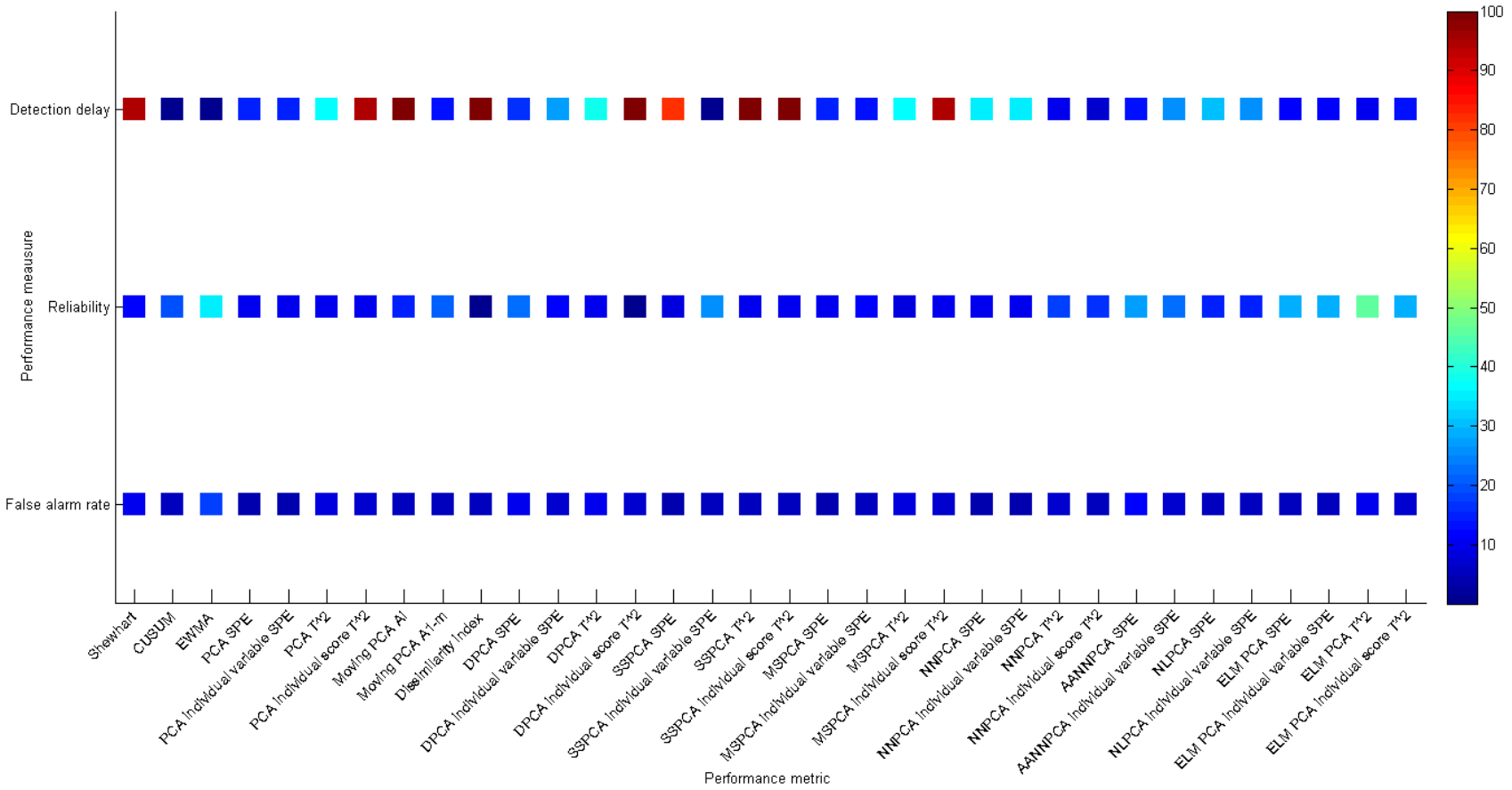

Figure 143: **Statistical data-based fault detection: primary mill – false alarm rates, reliability index and detection-delay at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

Next, a visual representation in the form of a CVA biplot (Figure 144) is made to determine if the primary mill performance drivers can be used to distinguish between the different recovery classes.

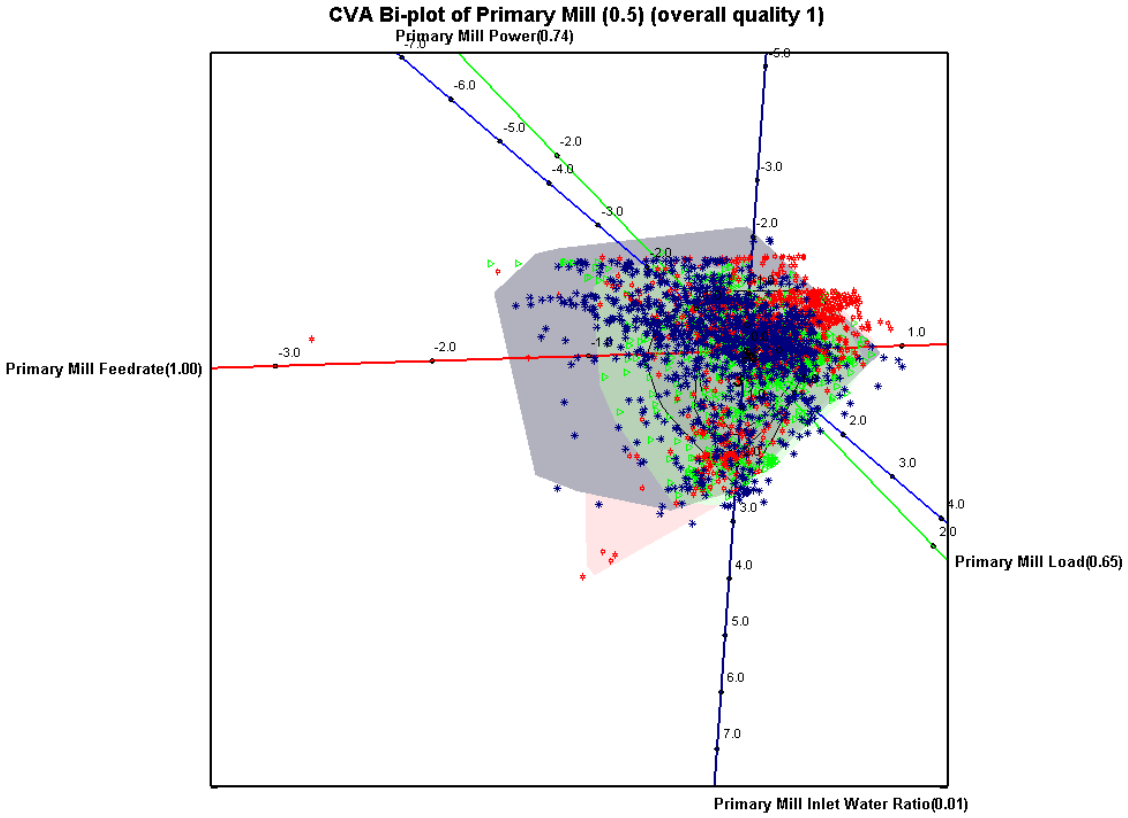

Figure 144: **CVA biplot: primary mill (class 1 – green; class 2 – red; class 3 – blue) at a confidence level of 0.5**

From the CVA biplot it is evident that the classes are not separated at all, with the alpha bags overlapping each other completely even at an alpha value of 0.5. Furthermore, since the primary mill inlet water ratio variable axis has a very low predictivity, significant and reliable conclusions can only be drawn with regards to the discriminatory ability of the primary mill feed rate, power and load. It can further be seen from the CVA biplot that there is a very high degree of correlation between the primary mill power load variables, as is evident from the small angle between the biplot axes for these measurements. This supports the structure of the fundamentally derived milling process causality map.

As before, the CVA biplot analysis is supported by the variable importance analysis. With no visual differentiation between the classes using the CVA biplot analysis, the results from the variable importance analysis become crucial to the investigation. From the variable importance analysis on average over all 3 classes (Figure 145) the main differentiators between the recovery classes are the primary mill inlet water

Stellenbosch University http://scholar.sun.ac.za

ratio, primary mill feed rate and primary mill power. Moving only from class 1 to class 2 (Figure 146), this reduces to only the primary mill inlet water ratio and the primary mill feed rate being significant drivers with the primary mill inlet water ratio and the primary mill power being important drivers when moving from class 2 to class 3 (Figure 147). Of these drivers, the primary mill inlet water ratio was consistently an important driver, with the primary mill load (possibly due to its high correlation to the primary mill power) never being highlighted as being an important driver.

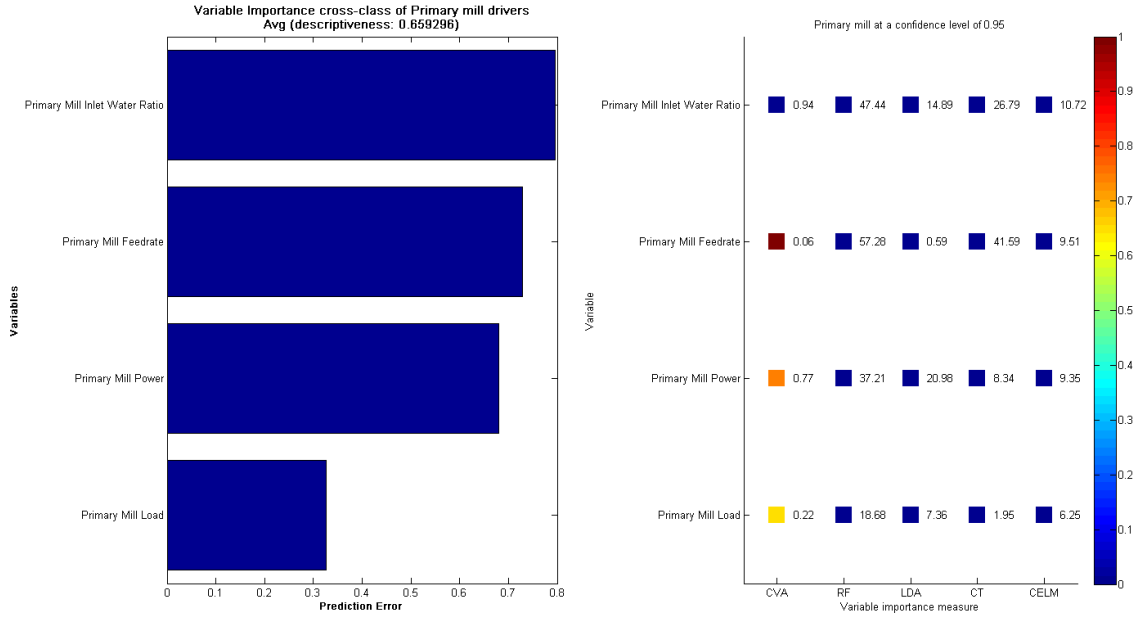

Figure 145: **Variable importance class 1-2-3: primary mill at a confidence level of 0.95**

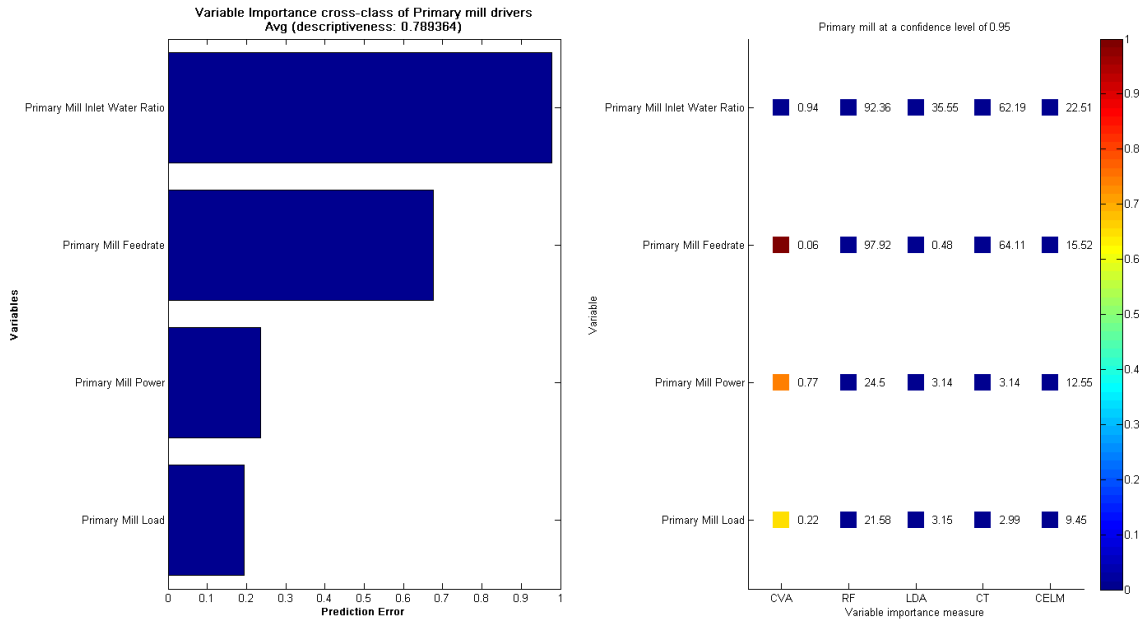

Figure 146: **Variable importance class 1-2: primary mill at a confidence level of 0.95**

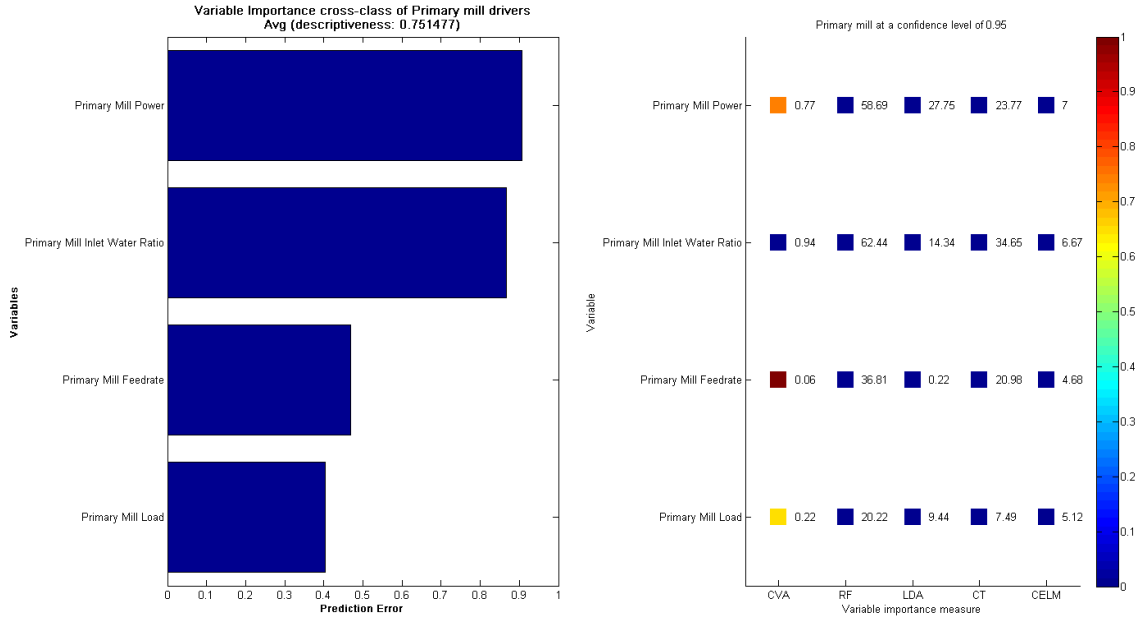

Figure 147: **Variable importance class 2-3: primary mill at a confidence level of 0.95**

Stellenbosch University http://scholar.sun.ac.za

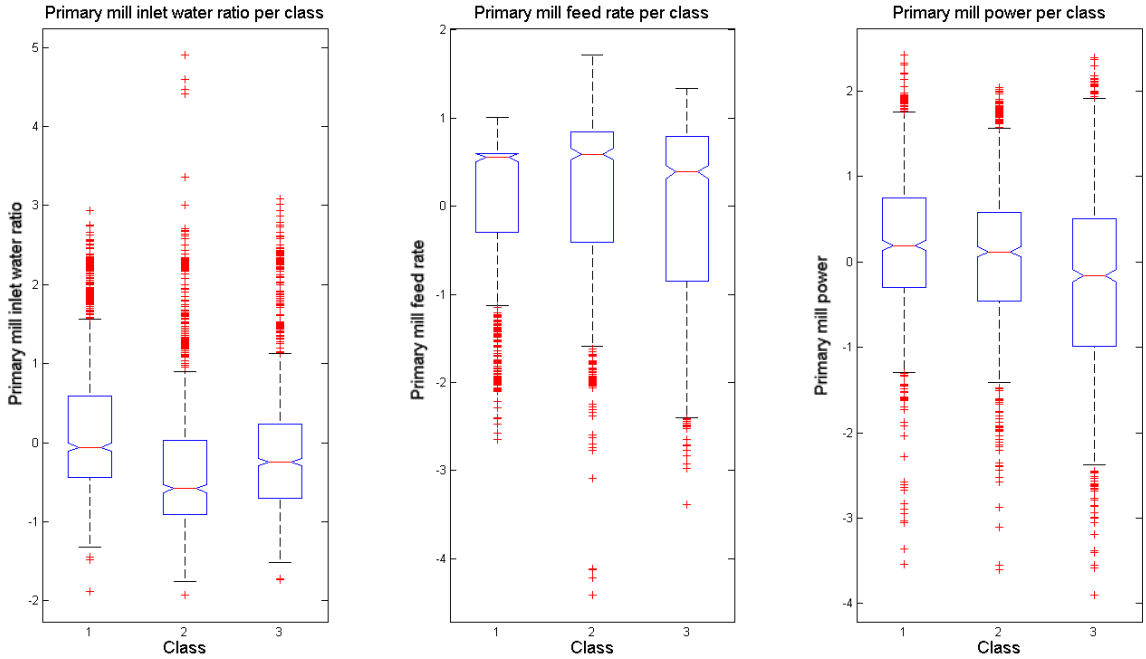

Figure 148: **Median significance: primary mill inlet water ratio, primary mill feed rate and primary mill power**

Subjecting the 3 primary mill inlet water ratio, primary mill feed rate and primary mill power classes of data to the median significance (Figure 148) and one-way ANOVA analysis, it was found that there was a significant difference between the classes for all these variables at a significance level of less than 0.05. Although the CVA biplot was unable to separate the data classes (using the given variables), it can still be concluded that there is a good link between the recovery classes and the primary mill operating variables. Although the changes in the primary mill operating variables do not visually seem to be significant, they were in fact probably just large enough to cause the required changes in the chrome classification cyclone operating variables that resulted in the chrome classification cyclone underperforming.

Although a good correlation has been found thus far between the change points detected in the recovery variable and potential important variables related to the chrome classification performance, it would be interesting to see if similar change points, coinciding with the recovery change points, exist in the important variables. Similarly, it would be interesting to see if variables that have been shown to be important for specific class transitions show more distinct change points for those class transitions compared to others.

For the primary mill inlet water ratio variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), none of the significant change points detected in the recovery variable are noticeable in the change point detection results (Figure 149). This is a very similar result to what was found for change point detection performed on the higher frequency

Stellenbosch University http://scholar.sun.ac.za

chrome classification data, and once again can, in part, explain the significant overlap that was present in the CVA biplot (Figure 144). Two very distinct and significant change points were identified at around sample 1070 (by the SSA, ELM SSA and nearest-neighbours CUSUM techniques) and sample 1385 (by the SSA and ELM SSA techniques). However, both of these relate rather to upset conditions in the process and not state changes. Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

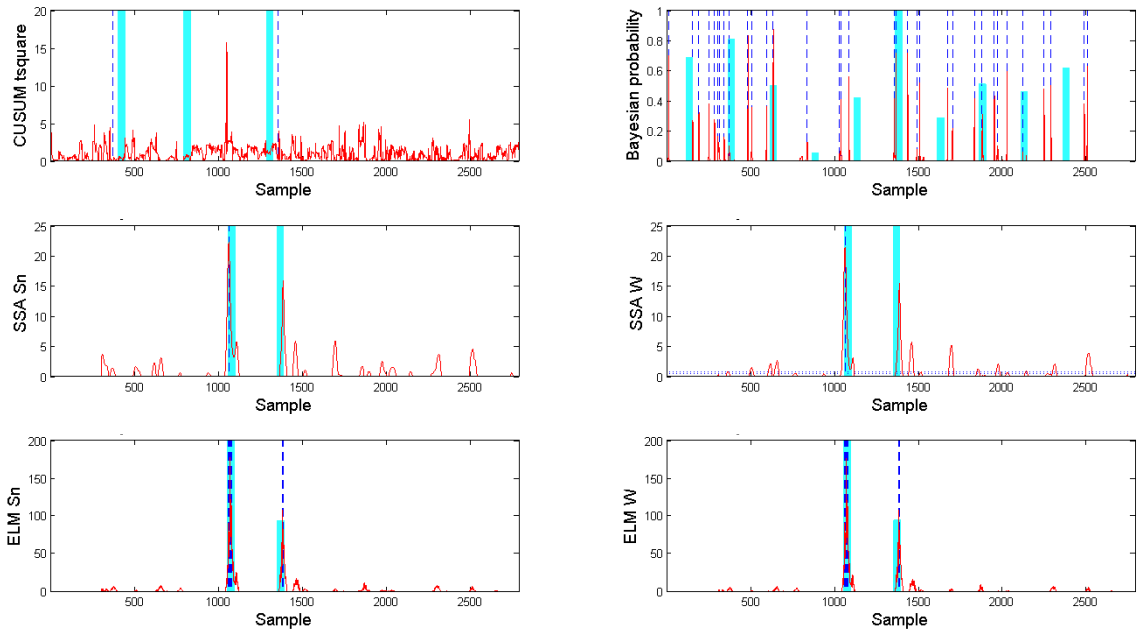

Figure 149: **Change point detection: primary mill inlet water ratio at a confidence level of 0.99**

For the primary mill feed rate variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), again none of the significant change points detected in the recovery variable are noticeable in the change point detection results (Figure 150). This is again a very similar result to what was found for change point detection performed on the higher frequency chrome classification data, and once again can, in part, explain the significant overlap that was present in the CVA biplot (Figure 144). All distinct and significant change points that were identified were found to relate rather to upset conditions in the process and not state changes. Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

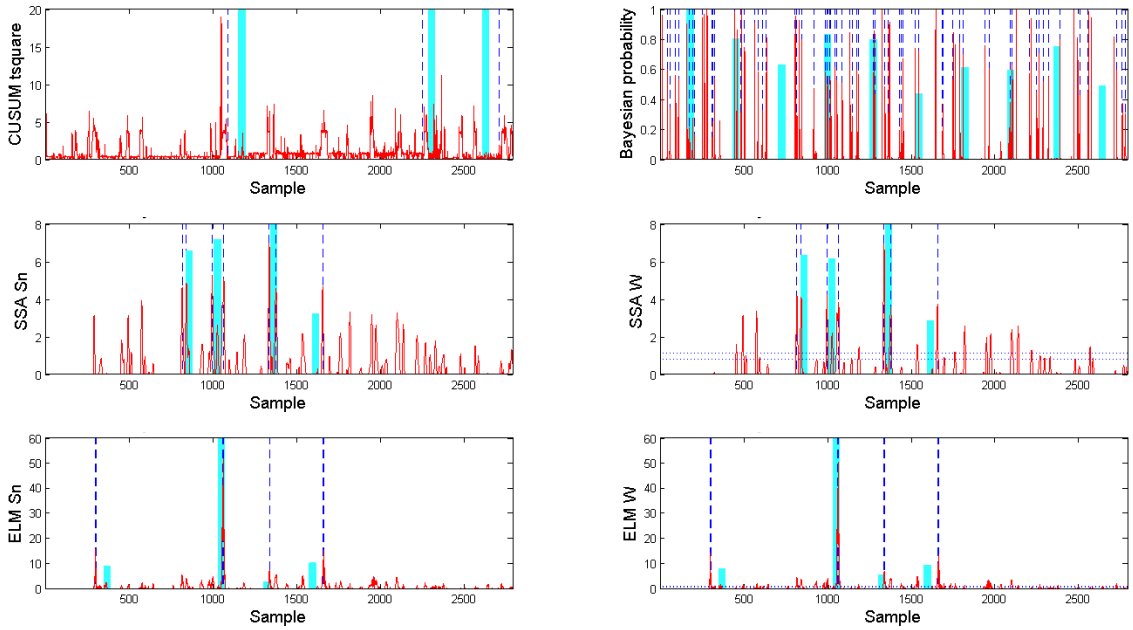

Figure 150: **Change point detection: primary mill feed rate at a confidence level of 0.99**

For the primary mill power variable (autocorrelated, not normally distributed, not exponentially distributed, SSA suggested change point detection technique), again none of the significant change points detected in the recovery variable are noticeable in the change point detection results (Figure 151). This is again a very similar result to what was found for change point detection performed on the higher frequency chrome classification data, and once again can, in part, explain the significant overlap that was present in the CVA biplot (Figure 144). All distinct and significant change points that were identified were found to relate rather to upset conditions in the process and not state changes. Other significant change points were also detected in the data, and should the root cause of the decrease in recovery prove difficult to determine, these change points should be investigated.

Stellenbosch University http://scholar.sun.ac.za

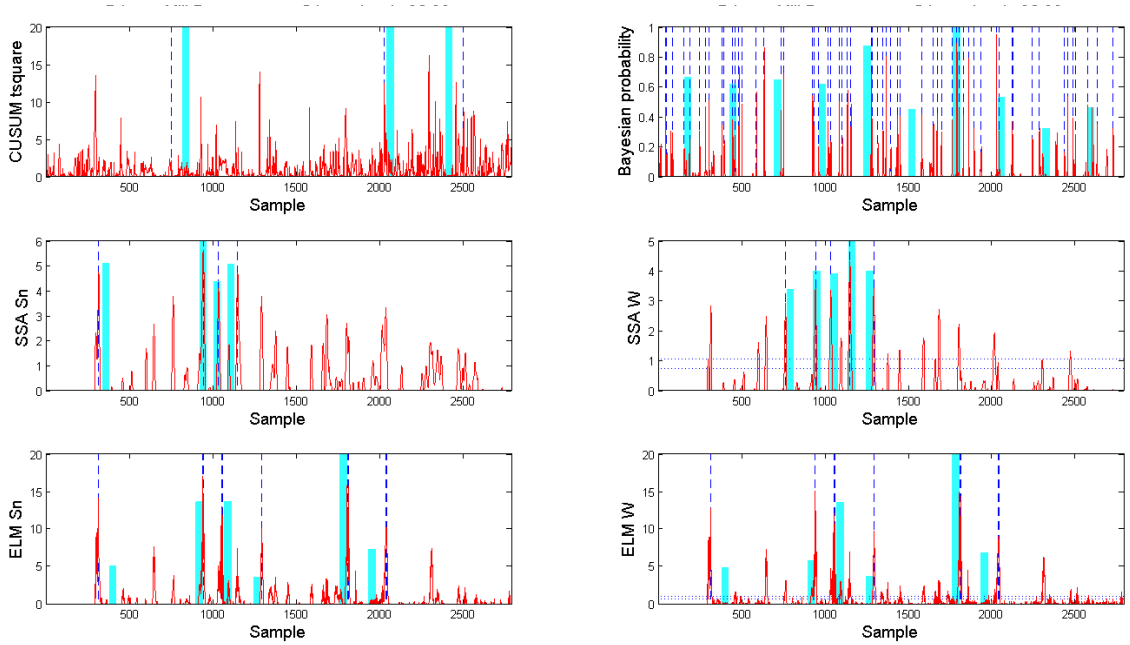

Figure 151: **Change point detection: primary mill power at a confidence level of 0.99**

The significance of the changes in the primary mill, especially the primary mill power and load, can be better visualised using higher frequency data. OPM graphs, in the form of mill operating curve plots, can also be used to estimate milling performance. From the mill operating curve plots for the primary mill (Figure 152) it can be seen that from class 1 to class 2 the mill performance moves down the mill operating curve, with both the mill power and mill load decreasing on average between the classes. In contrast to this, from class 2 to class 3 the mill performance not only moves to a new curve in the operating space, but it also moves further down this curve, again with both the mill power and load decreasing on average between the classes.

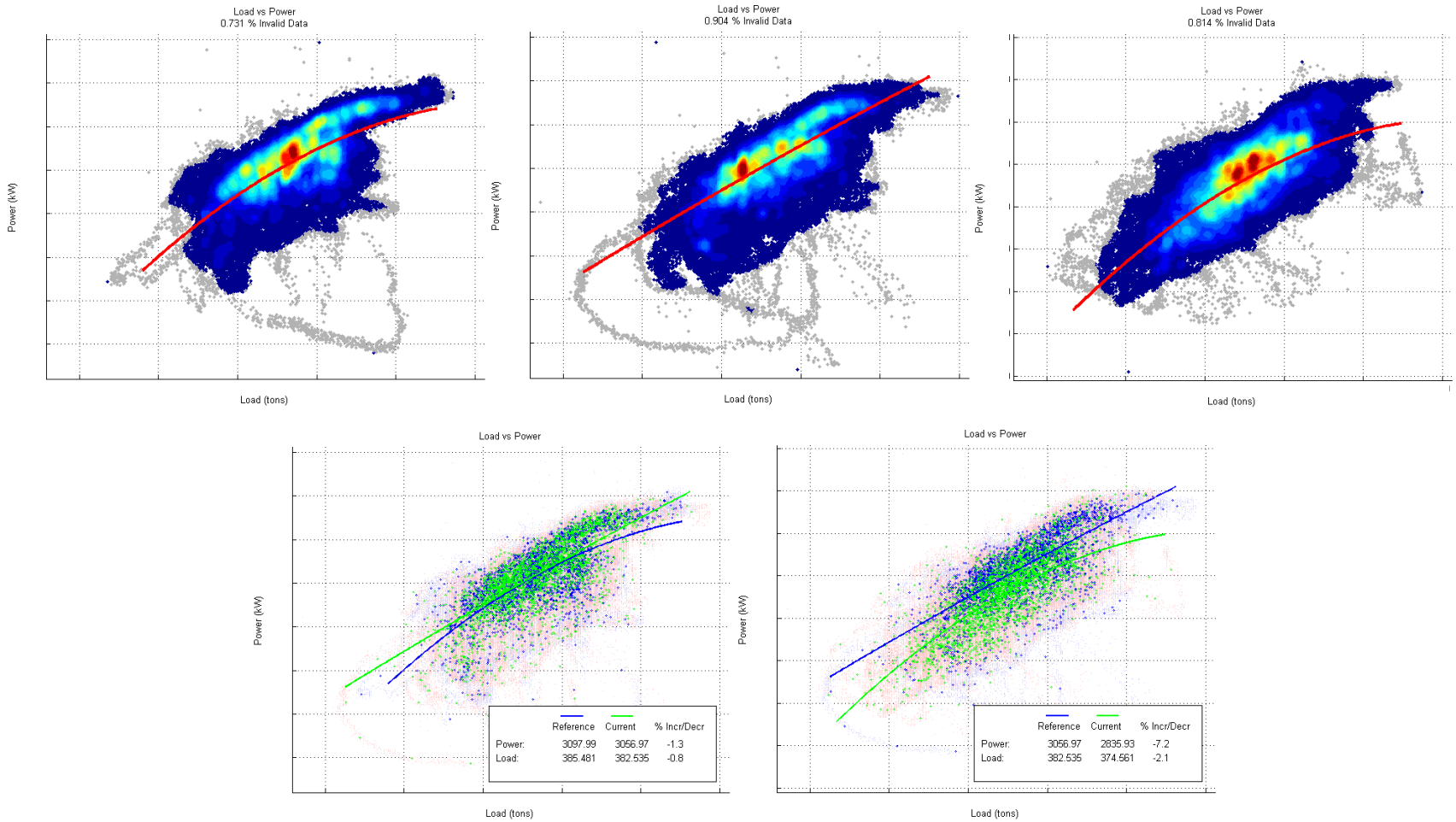

Figure 152: **OPM report: mill operating curve plots of primary mill (top left = class 1; top middle = class 2; top right = class3; bottom left: reference = class 1 & current = class 2; bottom right: reference = class 2 & current = class 3)**

Stellenbosch University http://scholar.sun.ac.za

At this stage of the analysis it is evident that the significant changes in the primary mill input variables (inlet water ratio and feed rate), causing both a consistent decrease in both the primary mill power and the primary mill load over all 3 data classes, resulted in a change in the feed to the chrome classification cyclone, causing the chrome classification cyclone to underperform, and ultimately contributing to the decrease in the recovery.

7.1.10 Summary

From the preceding analysis, the following can be summarised:

- Event: Recovery
  - A steady decline was visually noted in one of the main KPIs of the process: the recovery of the valuable mineral content.
  - Prior to a detailed analysis, all data was subjected to data validation and gross error detection, ensuring the quality/trustworthiness of the data.
  - Change point detection was performed on the recovery variable, with two significant change points being identified.
  - All data was subsequently divided into three classes (with the recovery variable consistently decreasing over all classes):
    - Class 1: reference data – for the period 01/10/2007 – 08/11/2007
    - Class 2: changeover data – for the period 08/11/2007 – 13/12/2007
    - Class 3: fault data for the – period 13/12/2007 – 01/02/2008
- Drivers: Recovery
  - The drivers of the shift in recovery were identified as mass pull (decreasing), final tail grade (increasing), final tail grind (coarsening) and plant feed head grade (decreasing).
  - Change points detected in these driver variables include:
    - Mass pull – 1st change point
    - Final tail grade – both change points
    - Plant feed head grade – 1st change point
    - Final tail grind – 1st change point – significant other change points detected – suspected to be due to data being more variable (noisy), frequently changing from day to day
  - Whereas the plant feed head grade measurement lies at the input side of the concentrator process and therefore a known disturbance root cause, the mass pull variable needed to be analysed as part of the flotation performance. The coarser final tail grind alluded to the fact that either the silica circuit or the chrome circuit was being overloaded and not performing optimally.
- Drivers: Final tail grade and grind
  - The drivers of the shift in the final tail grade and grind were identified as both, to a larger extent, the silica circuit tails grade (increasing) and grind (coarsening) and, to a much lesser extent, the chrome circuit tails grade (not significant) and grind (not significant).
  - Change points detected in these driver variables include:

Stellenbosch University http://scholar.sun.ac.za

        - Silica circuit tails grade – 1st change point – significant other change points detected
        - Chrome circuit tails grade – none – significant other change points detected
        - Silica circuit tails grind – none – significant other change points detected
        - Chrome circuit tails grind – both change points – significant other change points detected
- Drivers: Silica circuit grind
    - The drivers of the shift in the silica circuit grind were identified as the silica mill 01 power (increasing), the silica mill 02 power (increasing) and the primary flotation grind (not significant).
    - Change points detected in these driver variables include:
        - Silica mill 01 power – 2nd change point – significant other change points detected
        - Silica mill 02 power – none – significant other change points detected
- Drivers: Chrome circuit grade and grind
    - From the chrome circuit flotation performance analysis it was found that the material in the flash flotation cells had to have become less dense (having a larger water component).
    - From the chrome circuit milling performance analysis it was found that there must have been an increased feed rate (again higher water content) to the chrome mills.
    - Findings from both the silica circuit (increased solids feed rate) and the chrome circuit (increased water feed rate) indicates a decrease in the chrome classification performance/efficiency.
    - From the chrome classification cyclone performance it is evident that the mass split increased consistently over all the classes.
    - Change points were detected in chrome classification mass split in proximity to both the recovery change points.
- Drivers: Chrome classification
    - The drivers of the shift in the chrome classification mass split were identified as, to a larger extent, the chrome classification cyclone pressure (increasing) and feed flow rate (initially increasing) and, to lesser extent, the chrome classification cyclone feed density (decreasing).
    - Only significant other change points were detected, none related to the recovery change points.
    - Findings from the chrome classification indicate a shift in the primary mill operation.
- Drivers: Primary mill
    - The drivers of the shift in the primary mill operation were identified as the primary mill inlet water ratio (initially decreasing), primary mill feed rate (initially increasing) and primary mill power (decreasing).
    - Only significant other change points were detected, none related to the recovery change points.

Stellenbosch University http://scholar.sun.ac.za

- Although the changes in the primary mill operating variables did not visually seem to be significant, they were in fact probably just large enough to cause the required changes in the chrome classification cyclone operating variables that resulted in the chrome classification cyclone underperforming.

As the analyses progressed, moving further down the concentrator process causality map, the data classes seem to become more misaligned. This misalignment, as was evident by the decrease in reliability of the fault detection performance metrics and the increase in overlap between the CVA biplot classes, could be a result of the interacting nature of the process or even the effect of operator actions. It is, however, indicative of the fact that different drivers, at different times, had different effects on the process, however, they all contributed and could be traced back through the process as significant drivers of change in the process. No single variable was therefore solely responsible for the decrease in recovery, but a combination of events, with some being more significant than others. Also, some changes that occurred in the process did not necessarily contribute to the change in the process, but contributed to sustaining the decrease in recovery. It is further evident that since the classes did not always align perfectly, it would not have been possible to simply review the process trends to determine the root causes of the decrease in recovery.

The value of the standard OPM reports was also highlighted in enhancing the analyses. For the fault detection, the (multivariate, non-linear) extreme learning machine performance metrics together with the auto-associative neural networks performance metrics proved to be the most reliable throughout the analysis. The (univariate) CUSUM performance metric also proved exceptionally reliable, however, at the cost of having a high false alarm rate and a sometimes large detection-delay. It was shown that the techniques used weren't perfect, with the CVA biplot analysis and the variable importance analysis sometimes contradicting each other and the change point detection analysis seemingly giving more reliable results when applied to lower frequency data (more filtered, less noisy). It is speculated that for the higher frequency data the fact that the process data is not naturally stable manifested itself through the increased amount of change points detected. Furthermore, the change point detection algorithms are geared towards detecting any change in the data and not only process state changes. This in turn highlighted the importance of having a sound understanding of the process under investigation and applying this knowledge effectively as part of the analyses.

It was shown that the application of process causality maps greatly simplified the challenge of monitoring the process by reducing it to manageable portions. This allowed multiple, smaller, individual processes to be monitored at a low level, with the overall process being monitored at a higher level. Not only did this improve the ability of the techniques applied through a better focussed application, but the interpretability of the results also improved due to the reduction in complexity.

7.2 Benefit assessment

From a practical perspective for this case study, the plant under consideration gained value from the process performance monitoring in two key areas. Firstly, the developed fault detection models allowed

Stellenbosch University http://scholar.sun.ac.za

for the process to be continuously monitored for fault conditions, significantly reducing the potential future fault detection, identification and correction time and providing a structured approach to avoiding repeat failures. In addition to this, the automated OPM reports were also enhanced based on the findings of the case study, allowing for the improved monitoring and reporting of individual process unit operations on a daily basis.

Secondly, the case study afforded the plant personnel the valuable opportunity to learn from the data analysis. This was especially true for the more junior plant personnel where not only known cause and effect relationships were confirmed but also unknown relationships identified. Application of the process performance monitoring methodology increased the visibility into how problems are solved, what common root causes are and gave everyone concerned an appreciation of the complexity in finding the true root cause of a complex fault condition – providing tangible evidence of the cause and effect relationships allowed decision making to be based less on opinion and more on fact. To this end, the following key insights regarding the operation of the process were gained following the analysis:

- Seemingly insignificant changes in one part of the concentrator operation can have a significant knock-on effect in another part of the operation.
- The concentrator operation needs to be considered as a whole when determining the required process operating conditions; individual circuits should not be looked at in isolation.
- Classification is critical to any concentrator operation; getting it wrong can lead to key circuits being overloaded or underutilised.
- There are often overdesigned circuits with excess capacity within a concentrator operation; these typically act as buffers for when things go wrong, but this should not be depended upon.
- Having the right data available at the right time is critical in order to effectively monitor the performance of a process; this will also allow for informed decisions to be made about the process operation.

Unfortunately, the analysis was performed too long after the key event occurred. Corrective action was therefore not suggested as alternative action has already been taken by the time the analysis was completed. Although most of the data analysis algorithms can be applied in real time or near real time, practically, interaction from an experience user is currently still required in order to reason through the data analysis results. However, with the data analysis methodology in place, the process causality maps developed and access to the relevant process data streamlined, the turnaround time of similar, future root cause analyses have been significantly reduced, ensuring the timely implementation of suggested corrective action.

The methodology and associated infrastructure further allows more problems to be solved remotely, avoiding the costly unnecessary duplication of skills. Where skills duplication cannot be avoided, training time is significantly reduced owing to the supporting infrastructure.

Stellenbosch University http://scholar.sun.ac.za

8 A logistical perspective on the implementation of process monitoring systems

It is often the case that businesses do not have the required resources (both human and non-human) with which to effectively exploit all available information that is generated and stored. In order to make the most of these scarce resources, a well-defined implementation strategy is critical, especially if the goal is to achieve effective process performance monitoring. Key components of such a strategy typically include infrastructure, staff and skills, and system maintenance, often requiring major changes with regards to human resources, management, processes and company culture. In this context, Anglo American Platinum has had varying degrees of success with the different components of the proposed process performance monitoring methodology. Although the concept of the proposed methodology is sound, the implementation thereof within Anglo American Platinum still needs much work before it can be labelled a resounding success.

This chapter provides a review of key components as well as an evaluation of the practical issues with regards to the implementation of the developed process performance monitoring methodology for mineral processing plants.

8.1 Infrastructure

8.1.1 Software

Initial software design and development was guided by the findings of Miletic et al. (2004) for the development of an industrial data analysis application with long-term use as a central theme:

- System design – the difference between desktop analysis tools and a working on-line system should not be underestimated. Key design objectives typically include ease of system integration, maintenance, timely acquisition and storage of data, separation between data routing and model computations (allowing for portability), maximising the use of standard programming tools (minimising risk in long-term support), historical data playback and ease of access to both original process data and model results, providing maximum information on the minimum number of screens.
- Integration – must be done to ensure stable integration into existing system frameworks, requiring extensive testing and end-user training.
- Evaluation – systems evaluation should be done to allow the direct evaluation of the usefulness of the system. Typical performance indicators could include rates for computing and storing events, hours of system up-time and the number of likely events avoided through use of the system (cost saving).

Given the system design considerations, MATLAB (Mathworks) was selected as the software development platform. MATLAB is not only a dedicated, reliable computational environment for

Stellenbosch University http://scholar.sun.ac.za

engineering software development, but also makes available to the user a comprehensive collection of formal toolboxes, having optimised numerical processing and rich data visualisation features. MATLAB, through its Database Toolbox, also allowed for seamless integration to both the different process historians in use – InSQL (Wonderware) and PI (OSIsoft) – and the laboratory SQL (Microsoft) database. The developed software not only allowed for standardised automated data analysis consisting of data collection, validation, analysis and reporting, but also allowed for the reuse of the pre-processed data for manual ad hoc data analysis.

The process performance monitoring software was initially developed as a "proof of concept", covering only a fraction of the available sites and operating units within Anglo American Platinum. The "proof of concept" software, having no graphical user interface (GUI), was command line driven and expert operators were required to use the software. Over time, as the user base increased, the software functionality was extended to cater for new requirements and an increased analysis footprint. This placed a considerable amount of strain on the "proof of concept" code and required a complete rewrite of the software. Based on the learnings from the initial "proof of concept" code, the process performance monitoring software was redesigned and rewritten, not only making use of new features available in MATLAB to improve execution speed, reduce memory demand and improve scalability but to also allow for flexible customisation by the end-user or developer, deployment as a standalone executable and rapid deployment after code changes.

The selection of software development platform and subsequent developed process performance monitoring software has been deemed a success, especially considering its flexibility, expandability and integrability with other systems.

8.1.2 Hardware

Anglo American Platinum's operations are geographically quite distributed, being located in the North West and Limpopo Provinces of South Africa as well as Zimbabwe. This means that the data sources, plant historians and laboratory databases, are also quite distributed. Given this layout, a data analysis server per operation would be ideal. Unfortunately, this would require a large capital layout for the large amount of servers and software required as well as yearly reoccurring costs for maintenance and support. Furthermore, utilisation of the servers would be relatively low due to automated reports only being generated once a day, with ad hoc analyses only being performed infrequently.

In order to reduce cost and complexity, and also increase hardware utilisation, it was decided to house a limited amount of data analysis servers at a central location. The disadvantages of this solution was that not all reports could be generated simultaneously and that if there were network related problems the operations did not have access to their process performance monitoring reports, affecting availability of the system. Fortunately, the monitoring of all operations was not always equally important, allowing prioritisation of the different operations and their subsequent analysis in order of importance. Also,

Stellenbosch University http://scholar.sun.ac.za

network uptime for Anglo American Platinum can be considered high, resulting in high availability of the process performance monitoring system.

Recently, the system has also been moved from having a physical server per deployment to using virtual machines (VMs). This not only resulted in a smaller physical footprint for the process performance monitoring system through the use of more powerful servers that could be used to house multiple VMs, but also easier setup, configuration and maintenance of the system. Since the data analysis servers all have the same setup and configuration, a single VM needed to be created and configured and then simply rolled out as required. Availability of the system is also increased significantly through the fact that if anyone of the VMs fail, the master VM is simply restored, this being much faster than having to manually setup a new server from scratch.

The hardware selection and configuration has also been deemed a success, especially considering its cost, complexity and maintenance requirements compared to a completely distributed solution.

8.1.3 Techniques

Initial data analysis technique selection was also guided by the findings of Miletic et al. (2004), ensuring that the application of the selected techniques is feasible for meeting the business objectives:

- Data selection and preparation – variables should preferentially be selected for carrying useful information on an event or a quantity to be predicted rather than having a theoretical basis but not showing a strong signal. It is suggested that one initially select as many variables as possible for the analysis, after which the preliminary results should be used to screen the variables for improving the model with regard to accuracy and noise reduction based on process understanding. Data should be pre-processed by scaling, weighting, alignment for data collected at different rates or at different locations in the process, outlier detection, filtering to remove known drifts and selection to span the entire NOC region. For predictive models the data should span the operating region in a balanced way.
- Model development – modelling involves activities related to model type selection, algorithm selection and parameter estimation. Projection methods particularly have been found to lead to better numerical properties, providing the practitioner with model diagnosis tools and can improve the understanding of the data.
- Evaluation – off-line analysis is to be used for the evaluation of the applicability of using the selected techniques for meeting the particular objectives.

Data selection was supported through the use of process causality maps. This not only reduced the complex processes being monitored into manageable portions, but also improved the ability of the techniques applied through a better focussed application and the interpretability of the results through a reduction in complexity. With the data analysis results only being as reliable as the data on which it is based, data validation is probably the most important data preparation step. Checking for missing data,

Stellenbosch University http://scholar.sun.ac.za

not running data, high/low data, not updating data and rate of change exceeded data formed the core of the data validation, ensuring that only good quality data is used for process performance monitoring.

Appropriate data analysis techniques were selected based on the process performance monitoring objectives and data characteristics. The techniques covered a fair range of typical data characteristics with algorithm parameters being set to default values, based on literature and experience, where possible. Having multiple techniques available with which to estimate a particular outcome, e.g. variable importance, as well as being appreciative of the limitations of the various techniques ensured that a single technique never needed relying upon and allowed a combination of results to be interpreted within a very specific context. Data visualisation and the reporting interface proved critical to the success of the process performance monitoring system. It was found that although static reports presenting tables and graphs proved to be acceptable, ideally users required an interactive interface with which to further investigate results.

The applicability of the selected data analysis techniques need to be continuously evaluated to ensure their effectiveness in meeting the process performance monitoring objectives. New data analysis techniques and approaches also need to be regularly evaluated, not only in order to improve upon the current collection of data analysis techniques but also to allow new user requirements to be met.

The data analysis technique selection has also been deemed a success, especially considering the collection of available techniques and their applicability to the minerals processing data.

8.2 Staff and skills

Operational staff and their relevant skills are probably the most important factor responsible for a successful or failed implementation of process performance monitoring. Of these, a project sponsor or champion is probably the most critical, always ensuring that there is sufficient buy-in and momentum driving the concept. The project sponsor needs to have a strong understanding of the area of business where the process performance monitoring is being implemented as well as at least a cursory understanding of the technology involved. More importantly though, the project sponsor requires a degree of authority and respect within the organisation, ensuring that decisions are made and acted upon while simultaneously being a motivator and marketer for the acceptance of the technology.

For the implementation of the process performance monitoring methodology a three tiered approach was to be followed. At the bottom of the pyramid, tier 1, is the plant operational staff, responsible for the day-to-day monitoring of the processes. Tier 1 operators make use of their own custom trends and tables for decision making and troubleshooting as well as the standard automated process performance monitoring reports. Should a tier 1 operator become stuck with the analysis of a specific event, the event is escalated to tier 2 for assistance.

Stellenbosch University http://scholar.sun.ac.za

Tier 2 consists of operational staff that have been removed from various plants and relocated to a central monitoring centre. Tier 2 operators have been trained to use MATLAB as well as the advanced data analysis techniques contained within the process performance monitoring methodology. They are not only responsible for assisting tier 1 operators with requests but are also required to initiate independent data analysis tasks either based on events identified by fault detection or change point detection or longer term exploratory data analysis requests. Because of the metallurgical background of the tier 2 operators often not including any programming experience, Statistica (Statsoft), having a more Excel (Microsoft) type user interface was also introduced as an alternative data analysis platform. It was found that the tier 2 operators preferred using the Statistica software compared to the MATLAB developed software, and although containing an extensive collection of data analysis techniques the true potential of the process performance monitoring methodology was never reached by the tier 2 operators. Should a tier 2 operator become stuck with the analysis of a specific event, the event is escalated to tier 3 for assistance.

Tier 3 consists of a very limited number of operational staff assisting tier 2 operators with data analysis tasks, maintaining the systems on which the process performance monitoring relies as well as ensuring continuous improvement of the methodology and its associated data analysis techniques. Whereas tier 3 operators have a production experience background, their main expertise includes advanced data analysis, software programming and information systems.

From an implementation perspective, the staff and skills component have been identified as the principal weakness with an opportunity for improvement. Not only have financial constraints been a challenge, but a high staff turnaround also hindered the success of the system. Adoption of the process performance monitoring concept combined with an improved interface providing access to the data analysis techniques on which the methodology rests are probably the most challenging aspects of this component of the implementation, requiring the most attention going forward.

8.3 System maintenance

8.3.1 Software

Software maintenance of the process performance monitoring system has been relatively straightforward. As mentioned previously, there has been a considerable amount of general framework improvements to the “proof of concept” code resulting in a complete rewrite of the software with the new code being significantly more modular and scalable. Additionally, maintenance consisted of software bug fixes as well as numerous improvements and new functionality related to the data collection, validation, analysis and reporting components.

The software maintenance has therefore been deemed a success, considering no serious setbacks or stumbling blocks having been experienced.

Stellenbosch University http://scholar.sun.ac.za

8.3.2 Configuration

Configuration maintenance has been a very labour intensive process. Initial configuration data had to be configured manually based on process flow sheets and data availability in the process historians. Whenever plant changes occurred or process historians were updated, the process performance monitoring configuration data also needed updating. Recently the software has been modified to also allow for the configuration data to be obtained from the advanced process control (APC) system or the AF (OSIsoft) system.

The configuration maintenance can currently be deemed a partial success, largely due to the successful integration with the APC and AF systems. Unfortunately, reliance on these systems for configuration data has only resulted in a shift in responsibility for populating the configuration data. Work still needs to be done to improve the integration of these systems with the supervisory control and data acquisition systems.

8.3.3 Models

Model maintenance has also been relatively straightforward to date. The standard automated process performance monitoring reports are based on a concept of comparing “current” data to “reference” (benchmark) data. For data selection, the “reference” data set is selected to span a time period of similar duration prior and adjacent to the “current” data set. Each time a report is generated, the “current” data set would for example span the last 12 hours, 24 hours, 7 days, etc. of process operation with the “reference” data spanning a similar period prior and adjacent to the “current” data period. For ad hoc process performance monitoring, models are developed as needed and not currently typically retained afterwards for continued use.

The model maintenance can currently also be deemed a partial success, specifically with regards to the concept of comparing “current” to “reference” data having simplified the problem. There is, however, a growing need for moving towards the concept of having true NOC and known fault condition data, allowing for process condition “fingerprints” to be developed and monitored.

8.4 The road ahead

From an implementation perspective the immediate priority currently is associated with the people and skills component and requires an increase in the adoption of the process performance monitoring concept combined with an improved interface providing access to the data analysis techniques on which the methodology rests. Ways of improving the adoption of the process performance monitoring concept include, but are not limited to, expanding the use of the system to benchmarking and identifying best in class as well as process optimisation through the identification of better NOC regions. Improving the user interface to allow for easier access to the data analysis techniques is being pursued through the development of a custom GUI-based MATLAB user interface as well as closer integration with the plant information systems, specifically the OSIsoft software suite.

Stellenbosch University http://scholar.sun.ac.za

Longer term goals include having the process performance monitoring methodology fully automated, drilling through the process causality maps based on fault detection, change point detection and variable importance analysis results, and making use of predictive models, not only analysing events after the fact but also predicting future events and what will be causing them. This would lead to predictive diagnostics where corrective actions would be suggested to keep the process within the NOC region. Additionally, with return on investment being of critical importance to company shareholders, the application of process causality maps for performance monitoring should also be extended upwards to the enterprise level with the process performance measures ultimately including financial measures and targets.

Lastly, chances are good that as time progresses interest in process performance monitoring will decline, support will decrease and regularity will become erratic (Gosselin and Ruel, 2007). To counter these effects and sustain the project, a champion should lead the project, workflows should be established to maintain employees' interest, gains should be quantified, diffused and explained, and successes should be built upon.

Stellenbosch University http://scholar.sun.ac.za

9 Conclusions and recommendations

Process performance monitoring encompasses a broad collection of techniques aimed at monitoring processes not only for fault conditions but also for improvement opportunities. The application of process performance monitoring within Anglo American Platinum has been identified as an opportunity not only for providing expert support to operations but to also identify opportunities for process improvement. To this end the objective of this study has been to propose and evaluate a methodical approach to plant-wide process performance monitoring for mineral processing plants based on the concept of integrating process causality maps with data-based systems.

Hierarchical system decomposition was used as a framework for deriving process causality maps from a process system, not only retaining process flow topology and connectivity information via structural decomposition, but also the use of causal and fundamental process and unit specific information for variable selection and results interpretation via functional decomposition. It was found that the functional decomposition of the process, based on the fundamental characteristics of the process units, into unit specific causality maps significantly contributed to simplifying the practical challenge of variable selection (Kiran et al., 2012). Unit specific causality maps ensure that process variables are grouped when it fundamentally makes sense according to the process units. This not only improved the ability of the solution in linking the observed process behaviour to expected fundamental behaviour, but also reduced the likelihood of fundamentally incorrect conclusions being drawn from the data. Combining the unit specific causality maps with the structurally decomposed process system, based on the process flow topology and connectivity information, and the various statistical data-based fault detection techniques, especially the visually powerful representations offered by CVA biplots, significantly contributed to the user friendliness of the solution (Kiran et al., 2012). The process causality map ensured that an abnormal event could be visually tracked throughout the process to its root cause. Not only was this useful in highlighting cause and effect relationships throughout the process, but it also served to reveal previously unknown relationships within the process.

With statistical data-based techniques having historically been well studied and successfully employed in industrial applications (Venkatasubramanian et al., 2003c) it was proposed to implement change point detection and variable importance analysis techniques as complementary fault detection and diagnosis approaches. Change point detection complements statistical data-based fault detection techniques in that it allows the detection of wanted and unwanted events while the process is in a state of either normal or abnormal process operation. Being able to detect events during normal process operation makes it ideal for opportunity discovery with the aim of process improvement while still allowing for fault detection during abnormal process operation. For events detected through change point detection, where conventional fault detection models can be deemed to be inaccurate, unsuitable or unreliable, variable importance analysis was proposed for event diagnosis. This combination of change point detection and variable importance analysis, implemented as complementary approaches to conventional fault detection and diagnosis method, was found to contribute to addressing the practical challenge of technique reliability for highly non-linear process (Kiran et al., 2012).

Stellenbosch University http://scholar.sun.ac.za

In addition to this, novel, ELM-based techniques have been developed and implemented for:

- Statistical data-based fault detection: ELM PCA (based on neural network PCA fault detection).
- Change point detection: ELM SSA (based on SSA change point detection).
- Variable importance analysis: ELM-with-bagging (based on trees-with-bagging variable importance).

The extreme learning machine algorithm improves upon the more traditional neural network models by having extremely fast learning speeds, being good at generalisation and very having few parameters that require setting (Huang et al., 2006). Following the initial success of the NNPCA performance measures (Chen and Liao, 2002), it was proposed to replace the neural network algorithm with the ELM algorithm. It was found that the ELM PCA derived performance measures typically outperformed the NNPCA derived performance measures, attributed to its good generalization performance. Based on this improvement in performance it was proposed to implement the ELM algorithm as a residual generation stage, with the aim of removing the non-linear and dynamic characteristics from the data, prior to SSA change point detection, effectively resulting in a non-linear version of the SSA change point detection technique. Due to the way in which the SSA change point detection algorithm requires the continual updating of its base and test data matrices, the use of more slowly trained conventional feedforward neural networks is not possible, especially when considering on-line applications. For the evaluation case studies it was found that the performance of the proposed ELM SSA change point detection technique was slightly better than the SSA change point detection technique on which it is based. Although it has been shown that neither neural network nor CART models have a clear advantage of one over the other when considering prediction accuracy (Razi and Athappilly, 2005), it was proposed to replace the CART algorithm in the trees-with-bagging variable importance analysis approach with the ELM algorithm for classification. The proposed ELM-with-bagging approach, however, only managed to perform similarly to the trees-with-bagging approach for variable importance analysis.

In this chapter conclusions are drawn and recommendations made regarding each of the goals that needed to be attained in order to realise the proposed methodology.

9.1 Conclusions

The definition of process causality maps through the integration of causality with hierarchical system decomposition – Chapter 2. General rules were derived for the construction of process causality maps. Process causality maps not only allow for process and metadata to be logically grouped but also for the definition of process performance measures which suit the fundamental drivers of the process units. It was evident that the construction of process causality maps required not only a sound knowledge of how the various process variables relate to one another, but also a good fundamental understanding of how the various process units operate and interact with each another.

A critical literature survey on data-based statistical process monitoring techniques for fault detection, *the development and implementation of an extreme learning machine (ELM) based fault detection technique* and *a comparative evaluation of the various techniques on simulated time series data, simulated process*

Stellenbosch University http://scholar.sun.ac.za

data and a benchmark chemical plant simulation – Chapter 3. From the literature survey it was evident that extensive research has been done in the field of fault detection, covering a wide range of data structures and characteristics such as univariate, multivariate, static, dynamic, linear and non-linear. In addition to existing techniques, a novel ELM PCA algorithm was developed and implemented for fault detection. The evaluation of the fault detection techniques confirmed that for data having different structures and characteristics, no single technique is effective in detecting all potential fault conditions. It was noted that even when the data characteristics violate the assumptions underpinning a fault detection technique, often the technique will prove to be very robust in this regard and still be able to detect many of the fault conditions presented to it. The univariate and basic multivariate techniques were found to be very effective in detecting basic fault conditions in all the case studies evaluated. More advanced non-linear and dynamic multivariate techniques were, however, required for detecting many of the complex fault conditions in the Tennessee Eastman process case study, where relative to the other performance measures, the ELM PCA algorithm performed exceptionally well. It was, therefore, suggested that multiple fault detection techniques be run in parallel and if any one of them detects a fault condition the event be investigated.

A review of selected data-based change point detection techniques for inclusion in the process performance monitoring methodology in the context of event detection, *the development and implementation of an ELM-based change point detection technique* and *the testing and comparison of the techniques for the detection of process event conditions in simulated time series data, simulated process data and a benchmark chemical plant simulation – Chapter 4.* In addition to established techniques, a novel ELM SSA change point detection technique was developed and implemented. It was found for the change point detection technique evaluation that there is no single technique that is effective in identifying all potential fault conditions. Although, based on the data characteristics, the SSA change point detection technique was identified as the most appropriate for the majority of the case studies, it was noted that all of the change point detection techniques, and especially the Bayesian change point detection technique, proved to be exceptionally robust when the data characteristics violate the assumptions underpinning a fault detection technique. The proposed ELM SSA change point detection technique performed slightly better than the SSA change point detection technique on which it is based, specifically with regards to being more effective at detecting more subtle changes in the data such as a slowly shifting mean in the data (drift).

When comparing the change point detection results to the statistical data-based fault detection results, it was found that the change point detection techniques are at least as capable at detecting fault conditions as the statistical data-based fault detection techniques are. With the advantage of change point detection techniques over statistical data-based fault detection techniques being that they can be used to not only identify changes in process behaviour when changing from normal to abnormal process operation, but also changes in process behaviour while being in a normal or abnormal process state, their results can be used to partition data into different process states that can subsequently be monitored using statistical data-based fault detection techniques. Whereas pre-processing of the data could be considered to remove potential autocorrelation, possibly improving the Bayesian change point detection technique

Stellenbosch University http://scholar.sun.ac.za

results, applying multivariate fault detection techniques to the data for dimensionality reduction and feature extraction could be used to improve especially the (univariate) nearest-neighbours CUSUM change point detection technique by allowing the detection of changes in the fault detection space. As with the fault detection techniques, it was suggested that multiple change point analysis techniques be run in parallel and their results be interpreted in conjunction with expert process knowledge.

A review of selected data-based variable importance analysis techniques for inclusion in the process performance monitoring methodology in the context of event diagnosis and *the evaluation*, *the development and implementation of an ELM-based variable importance analysis technique and comparative performance of the techniques with regards to the identification and interpretation of important variables related to process event conditions in simulated time series data, simulated process data and a benchmark chemical plant simulation – Chapter 5.* In addition to established techniques, a novel ELM-with-bagging variable importance analysis technique was developed and implemented. It was found that all the variable importance techniques performed very similarly, with the proposed ELM-with-bagging variable importance analysis technique being a viable alternative to the CART based trees-with-bagging variable importance analysis technique. No single technique was found to be effective at identifying the important variables for all potential fault conditions. For the CVA biplot technique to effectively be used as a variable importance measure, a fair bit of visual interpretation was required in addition to the axes predictivities and adequacy values. The visual inspection of the results did, however, prove invaluable as to gaining an understanding of the important variables for each fault condition. Since only variables being monitored can be identified as important, the importance of monitoring as many of the process variables as possible was highlighted. This does, however, have the drawback that it becomes increasingly difficult to identify truly important variables, with visualisation of the results also becoming extremely complex. This is particularly relevant to the CVA biplot representations that become very crowded and difficult to interpret.

Correlation in the data was also found to be a challenge for all the variable importance techniques. This was especially true for the Tennessee Eastman process case study where it was noted that an absolute "correct" answer wasn't always obtainable regarding variable importance – the techniques not being able to distinguish between association and causation. When comparing the variable importance analysis results to the PCA SPE variable contribution results, it was found that the variable importance analysis techniques were at least as capable at identifying important variables as the PCA SPE variable contribution techniques were. The advantage of the variable importance analysis techniques over the PCA SPE variable contribution techniques lies with the fact that they can be used not only for analysing potentially interesting process events detected by change point detection, when statistical data-based fault detection is unsuitable, but also when the statistical data-based fault detection models are inaccurate or the contribution plots unreliable. As with the statistical data-based fault detection and change point detection techniques, even when the data characteristics of a data set violated the assumptions of the variable importance analysis techniques, the techniques proved to be fairly robust and still able to perform reasonably well. Again it was suggested that multiple variable importance analysis techniques be run in parallel and their results be interpreted in conjunction with expert process knowledge.

Stellenbosch University http://scholar.sun.ac.za

The development of an analytical methodology to plant-wide process performance monitoring for mineral processing plants, integrating process causality maps and data-based methods and *the development of process causality maps for a PGM concentrator plant through the integration of causality with hierarchical system decomposition – Chapter 6.* A 5-step analytical methodology was developed comprising of problem identification, data collection, data pre-treatment, data analysis and implementation of outcomes. Process causality maps were defined for key minerals processing equipment (crushers, mills, cyclones and flotation) and a generic concentrator process, describing the process to be monitored both fundamentally and statistically. Integration of process causality maps with data-based methods allowed, as part of the methodology, for the construction of a view of events showing the interaction between different variables and process operations. This was found necessary not only for identifying the root cause of a problem but also for highlighting the cause and effect relationships throughout the process as well as revealing previously unknown relationships within the process.

Although developed with a concentrator mineral processing plant in mind, various components of the methodology and the data-based analytical techniques have already been successfully applied to both smelting and refining processes within the mineral processing industry. From this it is evident that the methodology forms a fairly generic framework which is supported by process causality maps and data-based analytical techniques. Whereas the data-based analytical techniques have direct application in other industries, process specific causality maps would, however, have to be developed as required by the relevant industry: be it petrochemical, paper and pulp, food and beverages, fertilisers, pharmaceuticals, organic and geopolymers, biochemical, energy & water, etc.

The evaluation of this methodology through a comprehensive real-world mineral processing case study – Chapter 7. The mineral processing plant used in the case study was a PGM concentrating operation consisting of three grinding circuits with their associated flotation circuits. The process performance monitoring methodology was successfully applied to the case study, correctly detecting the fault condition, identifying the root causes and presenting a valuable opportunity to learn more about the process operation. While tracing the root cause of the event through the process, it was found that different drivers, at different times, had different effects on the process. It was shown how the different drivers contributed to the event, with no single variable being solely responsible for the event. Due to the naturally unstable nature of the process being evaluated, the techniques used were occasionally found inadequate, especially when analysing high frequency data, highlighting the importance of having a sound understanding of the process under investigation and applying this knowledge effectively as part of the analyses. The application of process causality maps was found to not only save time during the analysis through their reusability, but also greatly simplified the challenge of monitoring the process, not only improving the ability of the techniques applied through a better focussed application, but also the interpretability of the results due to the reduction in complexity.

From a practical perspective, value was gained through the developed statistical data-based fault detection models, not only significantly reducing future fault detection times, but also providing a

Stellenbosch University http://scholar.sun.ac.za

structured approach to avoid repeat failures. Furthermore, the automated OPM reports were enhanced based on the outcome of the analysis, allowing for improved monitoring and reporting of individual process unit operations on a daily basis. The outcomes of the analysis, providing such tangible evidence of cause and effect relationships, also allowed decision making to be based less on opinion and more on fact. Plant personnel were also afforded the valuable opportunity to learn from the data analysis, giving everyone concerned an appreciation of the complexity in finding the true root cause of a complex fault condition. Key insights gained regarding the operation of the process include the following:

- Seemingly insignificant changes in one part of the concentrator operation can have a significant knock-on effect in another part of the operation.
- The concentrator operation needs to be considered as a whole when determining the required process operating conditions; individual circuits should not be looked at in isolation.
- Classification is critical to any concentrator operation; getting it wrong can lead to key circuits being overloaded or underutilised.
- There are often overdesigned circuits with excess capacity within a concentrator operation; these typically act as buffers for when things go wrong, but this should not be depended upon.
- Having the right data available at the right time is critical in order to effectively monitor the performance of a process; this will also allow for informed decisions to be made about the process operation.

From a financial perspective, the timeous detection and analysis of the unwanted event coupled with appropriate corrective action could have reduced the loss of valuable material to the final tailings stream by up to 10%. For the period under review, depending on market prices and currency exchange rates, this equates to a loss of tens of millions of rands of potential revenue.

A critical evaluation of the practical issues with regards to the implementation of the methodology – Chapter 8. From a software perspective, MATLAB proved to be an excellent software development platform, especially considering its numerical processing capability and rich data visualisation features. MATLAB allowed the process performance monitoring software to be developed with a high degree of flexibility, expandability and integrability with other systems. Although Anglo American Platinum's operations are geographically quite distributed, it was decided to implement a centralised process performance monitoring system, not only reducing cost and complexity, but also increasing hardware utilisation. This decision was in part made possible due to the high network availability within the company, ensuring a high availability for the process performance monitoring system. Furthermore, virtual machines were implemented, reducing both the physical footprint and required maintenance of the process performance monitoring system.

From a data analysis perspective, extensive data validation guaranteed that only reliable, good quality data was used. Similarly, the application of process causality maps ensured appropriate data selection, with the process performance monitoring objectives and data characteristics being used to guide the selection of data analysis techniques. Algorithm parameters for the selected techniques were set to default values, based on literature and experience, where possible with remarkable success.

Stellenbosch University http://scholar.sun.ac.za

Furthermore, having multiple techniques with which to estimate a particular outcome, as well as being appreciative of the limitations of the techniques ensured that a single technique never needed relying upon and allowed a combination of results to be interpreted within a very specific context.

Operational staff and their relevant skills were identified as probably the most important factor responsible for a successful or failed implementation of process performance monitoring. A three tiered approach was implemented consisting of plant operational staff, responsible for the day-to-day monitoring of the processes, centralised metallurgical staff, responsible for assisting with root cause analysis type investigations, and centralised analytical staff, responsible for maintaining the systems on which the process performance monitoring relies as well as ensuring continuous improvement of the methodology and its associated data analysis techniques. Unfortunately, this approach was unsuccessful. Consequently, operational staff and their relevant skills, adoption of the process performance monitoring concept and an improved interface, providing access to the data analysis techniques on which the methodology rests, are the factors requiring the most effort, in the immediate future, to ensure a successful process performance monitoring implementation.

Maintenance of the process performance monitoring system has had varying degrees of success. Software maintenance has been relatively simple and straightforward with no serious setbacks or stumbling blocks having been encountered. Model maintenance has also been relatively straightforward to date, currently being based on a simple concept of comparing “current” to “reference” data and not typically retaining models for continued use. There is, however, a growing need for moving towards the concept of having true NOC and know fault condition data, allowing for process condition “fingerprints” to be developed and monitored. In contrast to this, configuration maintenance has been a very labour intensive process. Fortunately, due to the successful integration of the process performance monitoring system with the APC and AF systems, this should change in the near future.

9.2 Recommendations

For the purposes of this study, only generic *process causality maps* have been developed. It is recommended that for analyses where unit operation parameters and objectives differ (e.g. ball mill versus fully autogenous mill) or where process measurements are unavailable, these generic maps be customised, producing more accurate representations of the processes under investigation. Including site representation (experience plant personnel) during the development or review of the process causality maps will not only further improve the relevancy of the maps through the inclusion of previously observed behaviour, but also improve buy-in from the plant personnel regarding the methodology being applied. Additionally, with return on investment being of critical importance to company shareholders, the application of process causality maps for performance monitoring should also be extended upwards to the enterprise level with the process performance measures ultimately including financial measures and targets.

Whereas the process causality maps for this study were generated manually, allowing for the inclusion of fundamental process knowledge, automation of this task is currently partly possible through the use of

Stellenbosch University http://scholar.sun.ac.za

process connectivity information. The capturing of process connectivity information can effectively occur either via process flow diagram representations (Jiang et al., 2009; Maurya et al., 2003), available through documentation or process control systems, or process data (Bauer et al., 2007), available through historical production data. Given the maturity of the industrial systems in use, the automatic extraction of process connectivity information from documentation and process control systems is feasible (Thambirajah et al., 2009). Practically, the key requirement for the extraction of information is the availability and access to the required metadata representing the process flow and its associated control systems and instrumentation. This will allow for the automated generation of a process connectivity map, indicating how different process units and measurement instruments are connected, to which manually constructed unit specific causality maps can be added. In turn, it is expected that the extraction of process connectivity information from process data could be used to validate and/or improve the unit specific causality maps. This has, however, not yet been explored and needs to be investigated.

Following the evaluation of the selected *statistical data-based fault detection, change point detection and variable importance analysis* techniques, it is evident that the current collection of techniques does not cater for all potential data characteristics or expected events. It is important to continually research and evaluate new data analysis techniques (e.g. change point detection algorithms applicable to data that is not normally or exponentially distributed), ensuring the effectiveness of the process performance monitoring system in meetings its objectives. Although not perfect, the extreme learning machine based fault detection, change point detection and variable importance analysis techniques proved to be effective. Combined with the fact that the extreme learning machine algorithm has been shown to be robust and computationally inexpensive, with no real tuning parameters, it is an ideal candidate as core algorithm for the data analysis techniques forming part of a process performance monitoring solution.

The proposed process performance monitoring *methodology* and associated collection of data analysis techniques are geared towards manually investigating events after they have occurred. This has been found to be very time consuming and labour intensive. It is recommended that future investigations into this topic include the automation of the proposed methodology, automatically analysing and finding the optimal path from effect to cause through the process causality maps. Another significant improvement would be the use of predictive models, not only allowing events to be analysed after the fact but allowing the prediction of future events and suggested corrective actions to keep the process within the NOC region.

The current focus for the *implementation* of the process performance monitoring methodology has been on a mineral processing concentrator operation. Within Anglo American Platinum considerable benefits are still to be had by extending the implementation footprint to the other mineral processing operations: smelters and refineries. However, it has been highlighted that adoption of the process performance monitoring concept is still a major challenge, with only limited success having been achieved to date. It is suggested that this be addressed through an expanded application footprint focussing on current issues within the organisation including benchmarking, identifying best in class and process optimisation through the identification of better NOC regions.

Stellenbosch University http://scholar.sun.ac.za

Finally, chances are good that as time progresses interest in process performance monitoring will decline, support will decrease and regularity will become erratic (Gosselin and Ruel, 2007). To counter these effects and sustain the project, a champion should lead the project, workflows should be established to maintain employees' interest, gains should be quantified, diffused and explained, and successes should be built upon.

Stellenbosch University http://scholar.sun.ac.za

References

Abarbanel, H.D.I., 1996. *Analysis of Observed Chaotic Data*. New York: Springer-Verlag.

Abarbanel, H.D.I., Masuda, N., Rabinovich, M.I. and Tumer, E., 2001. Distribution of mutual information. *Physics Letters A*, 281, 368-373.

Adams, R.P. and KacKay, D.J.C., 2006. *Bayesian online changepoint detection* [online]. Cavendish Laboratory, University of Cambridge. Available at: http://www.cs.toronto.edu/~rpa/papers/rpa-changepoint.pdf [Accessed November 20, 2012].

Ahn, S.J., Lee, C.J., Jung, Y., Han, C., Yoon, E.S. and Lee, G., 2008. Fault diagnosis of the multi-stage flash desalination process based on signed digraph and dynamic partial least square. *Desalination*, 228, 68-83.

Aldrich, C., 2002. *Exploratory analysis of metallurgical process data with neural networks and related methods, Volume 1.* Amsterdam: Elsevier Science B.V.

Aldrich, C., Gardner, S. and Le Roux, N.J., 2004. Monitoring of metallurgical process plants by using biplots. *AIChE Journal*, 50 (9), 2167-2186.

Anderson, K.C., 2008. *A novel approach to Bayesian online change point detection* [online]. Department of Computer Science, University of Colorado Boulder. Available at: http://hdl.handle.net/10971/264 [Accessed November 20, 2012].

Archer, J.K. and Kimes, R.V., 2008. Empirical characterisation of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52, 2249-2260.

Atmanspacher, H., Scheingraber, H. and Voges, W., 1988. Global scaling properties of a chaotic attractor reconstructed from experimental data. *Physical Review A*, 37 (4), 1314-1322.

Auret, L. And Aldrich, C., 2010. Change point detection in time series data with random forests. *Control Engineering Practice*, 18, 990-1002.

Auret, L. And Aldrich, C., 2011. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105, 157-170.

Bakshi, B.R., 1998. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal*, 44 (7), 1596-1610.

Barry, D. And Hartigan, J.A., 1992. Product partition models for change point problems. *The Annals of Statistics*, 20 (1), 260-279.

Stellenbosch University http://scholar.sun.ac.za

Basila, M., Stefanek, G. and Cinar, A., 1990. A model-object based supervisory expert systems for fault tolerant chemical reactor control. *Computers and Chemical Engineering*, 14 (4-5), 551-560.

Bauer, M., Cox, J.W., Caveness, M.H., Downs, J.J. and Thornhill, N.F., 2007. Finding the Direction of Disturbance Propagation in a Chemical Process Using Transfer Entropy. *IEEE Transactions on Control Systems Technology*, 15(1), 12-21.

Bauer, M. and Thornhill, N.F., 2008. A practical method for identifying the propagation path of plant-wide disturbances. *Journal of Process Control*, 18, 707-719.

Becraft, W. and Lee, P., 1993. An integrated neural network/expert system approach for fault diagnosis. *Computers and Chemical Engineering*, 17 (10), 1001-1014.

Bie, L. and Wang, X., 2009. Fault Detection and Diagnosis of Continuous Process Based on Multiblock Principal Component Analysis. *In: International Conference on Computer Engineering and Technology,* Singapore 22-24 January 2009. IEEE Computer Society, 200-204.

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M. and Schröder, J., 2006. *Diagnosis and Fault-Tolerant Control.* Berlin: Springer-Verlag.

Born, M., 1949. *Natural Philosophy of Cause and Chance.* Oxford, UK: Clarendon.

Breiman, L., 1996. Bagging Predictors. *Machine Learning*, 24, 123-140.

Breiman, L., 1998. Arcing classifiers (with discussion). *Annals of Statistics*, 29, 801-849.

Breiman, L., 2001. Random forests. *Machine Learning*, 45, 5-32.

Breiman, L., 2001. Statistical Modelling: The Two Cultures. *Statistical Science*, 16 (3), 199-231.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Brown, D.E., Corruble, V. and Pittard, C.L., 1993. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems. *Pattern Recognition*, 26 (6), 953-961.

Bushveld complex, 2011 photograph. Available at: http://angloplatinum.investoreports.com/angloplatinum_iar_2011/downloads/angloplatinum_iar_2011.pdf [Accessed October 21, 2013].

Stellenbosch University http://scholar.sun.ac.za

Castro, J.J. and Doyle III, F.J., 2004. A pulp mill benchmark problem for control: problem description. *Journal of Process Control*, 14, 17-29.

Chang, C.C. and Yu, C.C., 1990. On-line fault diagnosis using the signed directed graph. *Industrial and Engineering Chemistry Research*, 29 (7), 1290-1299.

Chen, G. and McAvoy, T.J., 1998. Predictive on-line monitoring of continuous processes. *Journal of Process Control*, 8 (5-6), 409-420.

Chen, J. and Liao, C.M., 2002. Dynamic process fault monitoring based on neural network and PCA. *Journal of Process Control*, 12, 277-289.

Chen, L.W. and Modarres, M., 1992. Hierarchical decision process for fault administration. *Computer in Chemical Engineering*, 16 (5), 425-448.

Chen, R., Dave, K., McAvoy, T.J., and Luyben, M., 2003. A nonlinear dynamic model of a vinyl acetate process. *Industrial Engineering and Chemical Research*, 42 (20), 4478-4487.

Cheung, J.T. and Stephanopoulos, G., 1990. Representation of process trends part I: A formal representation framework. *Computers and Chemical Engineering* 14 (4-5), 495-510.

Chiang, L.H., Kotanchek, M.E. and Kordon, A.K., 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Computers and Chemical Engineering*, 28, 1389-1401.

Chiang, L.H., Russell, E.L. and Braatz, R.D., 2001. *Fault detection and diagnosis in industrial systems.* New York: Springer-Verlag.

Cho, J.-H., Lee, J.-M., Choi, S.W., Lee, D. and Lee, I.-B., 2005. Fault identification for process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 60, 279-288.

Choi, S.W., Changkyu, L., Jong-Min, L., Park, J.H. and Lee, I.B., 2005. Fault detection and identification of nonlinear processes based on kernel PCA. *Chemometrics and Intelligent Laboratory Systems*, 75, 55-67.

De Gooijer, J.G., 2006. Detecting change-points in multidimensional stochastic processes. *Computational Statistics & Data Analysis*, 51, 1892-1903.

Deng, X. and Tian, X., 2013. Nonlinear process fault pattern recognition using statistics kernel PCA similarity factor. *Neurocomputing*, 121, 298-308.

Stellenbosch University http://scholar.sun.ac.za

Detroja, K.P., Gudi, R.D. and Patwardhan, S.C., 2007. Plant-wide detection and diagnosis using correspondence analysis. *Control Engineering Practice*, 15, 1468-1483.

Ding, S., 2008. *Model-based Fault Diagnosis Techniques.* Berlin: Springer-Verlag.

Ding, S., Zhang, P., Ding, E., Engel, P. and Gui, W., 2011. A survey of the application of basic data-driven and model-based methods in process monitoring and fault diagnosis. *In:* S. Bittanti, A. Cenedese and S. Zampieri, *Proceedings of the 18th IFAC World Congress,* Milano, Italy 28 August – 2 September 2011. International Federation of Automatic Control, 12380-12388.

Dong, D. and McAvoy, T.J., 1996. Nonlinear principal component analysis based on principal curves and neural networks. *Computers in Chemical Engineering*, 20 (1), 65-78.

Dong, G., Chongguang, W., Beike, Z. and Xin, M., 2010. Signed Directed Graph and Qualitative Trend Analysis Based Fault Diagnosis in Chemical Industry. *Chinese Journal of Chemical Engineering*, 18 (2), 265-276.

Douglas, J.M., 1985. A hierarchical decision procedure for process synthesis. *AIChE Journal*, 31 (3), 353-362.

Downs, J.J. and Vogel, E.F., 1993. A plant-wide industrial process control problem. *Computers and Chemical Engineering*, 17 (3), 245-255.

Duan, P., Yang, F., Chen, T. and Shah, S.L., 2012. Detection of Direct Causality Based on Process Data. *In: American Control Conference,* Montreal 27-29 June 2009. IEEE Computer Society, 3522-3527.

Dundar, M., Krishnapuram, B., Bi, J. and Rao, R.B., 2007. Learning Classifiers when the training data is not IID. *In:* M. Veloso, *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India 6-12 January 2007. California: AAAI Press, 756-761.

Finch, F.E. and Kramer, M.A., 1987. Narrowing diagnostic focus using functional decomposition. *American Institute of Chemical Engineers Journal*, 34 (1), 130-140.

Fourie, S.H. and de Vaal, P., 2000. Advanced process monitoring using an on-line non-linear multiscale principal component analysis methodology. *Computers and Chemical Engineering*, 24, 755-760.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58 (3), 453-467.

Gardner, S., 2001. *Extensions of biplot methodology to discriminant analysis with applications of non-parametric principal components.* Thesis (PhD). University of Stellenbosch.

Stellenbosch University http://scholar.sun.ac.za

Gardner, S., Le Roux, N.J. and Aldrich, C., 2005. Process data visualisation with biplots. *Minerals Engineering*, 18, 955-968.

Gardner-Lubbe, S., Le Roux, N.J. and Gower, J.C., 2008. Measures of fit in principal component and canonical variate analyses. *Journal of Applied Statistics*, 35 (9), 947-965.

Genuer, R., Poggi, J-M. And Tuleau-Malot, C., 2010. Variable selection using random forests. *Pattern Recognition Letters*, 31, 2225-2236.

Gertler, J., 1998. *Fault Detection and Diagnosis in Engineering Systems*. New York, USA: Marcel Dekker Inc.

Glymour, C., 2003. Learning, prediction and causal Bayes nets. *TRENDS in Cognitive Sciences*, 7 (1), 43-48.

Gosselin, C. and Ruel, M., 2007. Advantages of monitoring the performance of industrial processes. *ISA Management Newsletter*, January 2007, 6-8.

Gower, J.C. and Hand, D.J., 1996. *Biplots*. London: Chapman & Hall.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37 (3), 424-438.

Grbovic, M., Li, W., Xu, P., Usadi, A.K., Song, L. and Vucetic, S., 2012. Decentralized fault detection and diagnosis via sparse PCA based decomposition and Maximum Entropy decision fusion. *Journal of Process Control*, 22, 738-750.

Hadler, K., Barbian, N., Cilliers, J.J., 2006. The relationship between the peak in air recovery and flotation bank performance. *Minerals Engineering*, 22 (5), 451-455.

Hadler, K., Smith, C.D. and Cilliers, J.J., 2010. Recovery vs. mass pull: The link to air recovery. *Mineral Engineering*, 23 (11-13), 994-1002.

Han, C., Shih, R. and Lee, L., 1994. Quantifying signed directed graphs with the fuzzy set for fault diagnosis resolution improvement. *Industrial and Engineering Chemistry Research*, 33 (8), 1943-1954.

Hastie, T., Tibshirani, R. And Frideman, J.H., 2009. *Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer-Verlag.

Stellenbosch University http://scholar.sun.ac.za

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M. and Bhattacharya, J., 2007. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441, 1-46.

Hodouin, D., MacGregor, J.F., Hou, M. and Franklin, M., 1993. Multivariate statistical analysis of mineral processing plant data. *CIM Bulletin*, 86 (975), 23-34.

Hotelling, H., 1947. Multivariate quality control illustrated by the testing of sample bombsights. *In:* O. Eisenhart, Selected techniques of statistical analysis. New York: Mcgraw-Hill, 113-184.

Hou, Q., Wang, L., Lu., N.Y., Jiang, B. and Lu, J.H., 2010. A FDD Method by Combining Transfer Entropy and Signed Digraph and its Application to Air Separation Unit. *In: The 11th International Conference on Control, Automation, Robotics and Vision,* Singapore 7-10 December 2010. IEEE Computer Society, 352-357.

Huang, G-B., 2003. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks*, 14(2), 274-281.

Huang, G-B., Zhu, Q-Y. and Siew, C-K., 2006. Extreme learning machine: Theory and applications. *Neurocomputing*, 70, 489-501.

Hwang, D-H. and Han, C., 1999. Real-time monitoring for a process with multiple operating modes. *Control Engineering Practice*, 7, 891-902.

Iri, M., Aoki, K., O'Shima, E. and Matsuyama, H., 1979. An algorithm for diagnosis of system failures in the chemical process. *Computers and Chemical Engineering*, 3 (1-4), 489-493.

Janusz, M. and Venkatasubramanian, V., 1991. Automatic generation of qualitative description of process trends for fault detection and diagnosis. *Engineering Applications of Artificial Intelligence*, 4 (5), 329-339.

Jemwa, G.T. and Aldrich, C., 2006. Kernel-based fault diagnosis on mineral processing plants. *Minerals Engineering*, 19, 1149-1162.

Jia, F., Martin, E.B. and Morris, A.J., 1998. Non-linear principal components analysis for process fault detection. *Computers in Chemical Engineering*, 22, S851-S854.

Jiang, H., Patwardhan, R. and Shah, S.L., 2009. Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix. *Journal of Process Control*, 19(8), 1347-1354.

Jiang, Q. and Yan, X., 2012. Chemical processes monitoring based on weighted principal component analysis and its application. *Chemometrics and Intelligent Laboratory Systems*, 119, 11-20.

Stellenbosch University http://scholar.sun.ac.za

Juricek, B.C., Seborg, D.E. and Larimore, W.E., 2001. Predictive monitoring for abnormal situation management. *Journal of Process Control*, 11, 111-128.

Kano, M., Hasebe, S. and Hashimoto, I., 2002. Statistical process monitoring based on dissimilarity of process data. *AIChE Journal*, 48 (6), 1231-1240.

Kano, M., Hasebe, S., Hashimoto, I. and Ohno, H., 2001. A new multivariate statistical process monitoring method using principal component analysis. *Computers and Chemical Engineering*, 25, 1103-1113.

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R. and Bakshi, B.R., 2002. Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem. *Computers and Chemical Engineering*, 26, 161-174.

Kharva, M., 2001. *Monitoring of froth systems using principal component analysis*. Thesis (MSc). University of Stellenbosch.

Kiran, K.L., Selvaraj, S. and Hua, J.L.C., 2012. Application of fault monitoring and diagnostic techniques and their challenges in petrochemical industries. *In: Preprints of the 8th IFAC Symposium on Advanced Control of Chemical Processes,* Singapore 10-13 July 2012. The International Federation of Automatic Control, 702-707.

Kotz, S. and Johnson, N.L., 1982. *Encyclopedia of Statistical Sciences*. New York: John Wiley and Sons.

Kourti, T. and MacGregor J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28, 3-21.

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37 (2), 233-243.

Kramer, M.A. and Palowitch, B.L., 1987. A rule based approach to fault diagnosis using signed directed graph. *American Institute of Chemical Engineers Journal*, 33 (7), 1067-1078.

Kruger, U., Zhou, Y. and Irwin, G.W., 2004. Improved principal component monitoring of large-scale processes. *Journal of Process Control*. 14, 879-888.

Ku, W., Storer, R.H. and Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Systems*, 30, 179-196.

Stellenbosch University http://scholar.sun.ac.za

Kugiumtzis, D., 1996. State space reconstruction parameters in the analysis of chaotic time series – the role of the time window length. *Physica D*, 95 (1), 13-28.

Kugiumtzis, D. and Christophersen, N., 1997. *State space reconstruction: method of delays vs. singular spectrum approach* [online]. Department of Informatics, University of Oslo. Available at: http://ftp.project.ifi.uio.no/publications/research-reports/DKugiumtzis-1.pdf [Accessed November 20, 2012].

Kumamoto, H., Ikechi, K., Inoue, K. and Henley, E.J., 1984. Application of expert system techniques to fault diagnosis. *The Chemical Engineering Journal*, 29 (1), 1-9.

Lee, J.-M., Yoo, C., Choi, S.W., Vanrollghem, W.A. and Lee, I.-B., 2004. Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science*, 59, 223-234.

Larsson, J.E., 1994. Diagnostic reasoning strategies for means-end models. *Automatica*, 30 (5), 775-787.

Lin, W., Qian, Y. and Li, X., 2000. Nonlinear dynamic principal component analysis for on-line process monitoring and diagnosis. *Computers and Chemical Engineering*, 24, 423-429.

Lind, M., 1991. Abstraction for modelling diagnostic strategies. *In: IFAC Workshop on Computer Software Structures Integrating AI/KBS Systems in Process Control,* Norway 29-30 May 1991.

Lowry, C.A., Woodall, W.H., Champ, C.W. and Rigdon, S.E., 1992. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34 (1), 46-53.

Lu, N., Jiang, B., Wang, L., Lu, J. and Chen, X., 2012. A Fault Prognosis Strategy Based on Time-Delayed Digraph Model and Principal Component Analysis. *Mathematical Problems in Engineering*, 2012, 1-17.

MacGregor, J.F., Jaeckle, C., Kiparissides, C. and Koutodi, M., 1994. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal*, 40, 826-828.

MacGregor, J.F. and Kourti, T., 1995. Statistical process control of multivariate processes. *Control Engineering Practice*, 3 (3), 403-414.

Matis, K.A., Gallios, G.P. and Kydros, K.A., 1993. Separation of fines by flotation techniques. *Separation Technology*, 3, 76-90.

Stellenbosch University http://scholar.sun.ac.za

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V., 2003. A Systematic Framework for the Development and Analysis of Signed Digraphs for Chemical Processes. 2. Control Loops and Flowsheet Analysis. *Industrial Engineering and Chemical Research*, 42 (20), 4811-4827.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V., 2004. Application of signed digraphs-based analysis for fault diagnosis of chemical process flowsheets. *Engineering Applications of Artificial Intelligence*, 17, 501-518.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V., 2005. Fault diagnosis by qualitative trend analysis of the principal components. *Chemical Engineering Research and Design*, 83 (A9), 1122-1132.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V., 2007a. A signed directed graph and qualitative trend analysis-based framework for incipient fault diagnosis. *Chemical Engineering Research and Design*, 85 (A10), 1407-1422.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V., 2007b. Fault diagnosis using dynamic trend analysis: A review and recent developments. *Engineering Applications of Artificial Intelligence*, 20, 133-146.

Maurya, M.R., Paritosh, P.K., Rengaswamy, R. and Venkatasubramanian, V., 2010. A framework for on-line trend extraction and fault diagnosis. *Engineering Applications of Artificial Intelligence*, 23, 950-960.

McAvoy, T.J. and Ye, N., 1994. Base control for the Tennessee Eastman Problem. *Computers and Chemical Engineering*, 18, 383-413.

Miletic, I., Quin, S., Dudzic, M., Vaculik, V. and Champagne, M., 2004. An industrial perspective on implementing on-line applications of multivariate statistics. *Journal of Process Control*, 14, 821-836.

Modarres, M., Cheon, S.W., 1999. Function-centered modelling of engineering systems using the goal tree-success tree technique and functional primitives. *Reliability Engineering and System Safety*, 64, 181-200.

Moskvina, V., 2001. *Application of the Singular-Spectrum Analysis for Change-Point Detection in Time Series.* Thesis (PhD), Cardiff University.

Moskvina, V. And Zhigljavsky, A., 2003. An algorithm based on singular spectrum analysis for change-point detection. *Communication in Statistics*, 32 (2), 319-352.

Nan, C., Khan, F. and Iqbal, M.T., 2008. Real-time fault diagnosis using knowledge-based expert systems. *Process Safety and Environmental Protection*, 86, 55-71.

Stellenbosch University http://scholar.sun.ac.za

Nandi, S., Badhe, Y., Lonari, J., Sridevi, U., Rao, B.S., Tambe, S.S. and Kulkarni, B.D., 2004. Hybrid process modelling and optimization strategies integrating neural networks/support vector regression and genetic algorithms: study of benzene isopropylation on Hbeta catalyst. *Chemical Engineering Journal*, 97, 115-129.

Napier-Munn, T.J., Morrell, S., Morrison, R.D. and Kojovic, T., 1999. *Mineral Comminution Circuits: Their Operation and Optimisation*. Australia: Julius Kruttschnitt Mineral Research Centre.

Nichols, J.M. and Nichols, J.D., 2001. Attractor reconstruction for non-linear systems: a methodological note. *Mathematical Biosciences*, 171, 21-32.

Nijhuis, A., de Jong, S. and Vandeginste, B.G.M., 1997. Multivariate statistical process control in chromatography. *Chemometrics and Intelligent Laboratory Systems*, 38, 51-62.

Nomikos, P. and MacGregor, J.F., 1995. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1), 41-59.

Ogunnaike, B.A. and Ray, W.H, 1994. *Process Dynamics, Modelling, and Control.* New York: Oxford University Press.

Olshausen, B.A., 2004. *Bayesian probability Theory* [online]. Redwood Center for Theoretical Neuroscience, University of California. Available at: http://redwood.berkeley.edu/bruno/npb163/bayes.pdf [Accessed November 20, 2012].

Page, E. S., 1954. Continuous Inspection Scheme. *Biometrika* 41 (1/2), 100–115.

Paquet, U., 2007. Empirical Bayesian change point detection [online]. Department of Process Engineering, University of Stellenbosch. Available at: http://www.ulrichpaquet.com/Notes/changepoints.pdf [Accessed November 20, 2012].

Patton, R., Frank, P. and Clark, R., 2000. *Issues of Fault Diagnosis for Dynamic Systems*. Berlin: Springer-Verlag.

Pearl, J., 2003. Causality: models, reasoning, and inference. *Econometric Theory*, 19, 675-685.

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P. and Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, 207, 304-318.

Prasad, P.R., Davis, J.F., Jirapinyo, Y., Josephson, J.R. and Bhalodia, M., 1998. Structuring diagnostic knowledge for large-scale process systems. *Computers in Chemical Engineering*, 22 (12), 1897-1905.

Stellenbosch University http://scholar.sun.ac.za

Questier, F., Put, R., Coomans, D., Walczak, B. And Vander Heyden, Y., 2005. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory System*, 76, 45-54.

Qian, Y., Li, X., Jiang, Y. and Wen, Y., 2003. An expert system for real-time fault diagnosis of complex chemical processes. *Expert Systems with Applications*, 24, 425-432.

Qin, S.J., Valle, S. and Piovoso, M.J., 2001. On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15, 715-742.

Qin, S., 2003. Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, 17, 480-502.

Qin, S., 2009. Data-driven fault detection and diagnosis for complex industrial processes. *In:* T. Escobet, V. Puig and B. Morcego, *7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes,* Barcelona, Spain 30 June – 3 July. Elsevier Science, 1115-1125.

Qin, S.J. and Li, W., 2001. Detection and identification of faulty sensors in dynamic processes. *American Institute of Chemical Engineering Journal*, 47 (7), 1581-1593.

Qingchao, J. and Xuefeng, Y., 2013. Statistical monitoring of Chemical Processes based on Sensitive Kernel Principal Components. *Chinese Journal of Chemical Engineering*, 21 (6), 633-643.

Rajamani, R.K. and Herbst, J.A., 1991. Optimal control of a ball mill grinding circuit: Part I Grinding circuit modelling and dynamic simulation. *Chemical Engineering Science*, 46 (3), 861-870.

Ramesh, T.S., Shum, S.K. and Davis, J.F., 1988). A structured framework for efficient problem-solving in diagnostic expert systems. *Computers and Chemical Engineering*, 9-10 (12), 891-902.

Rasmussen, J., 1985. The role of hierarchical knowledge representation in decision making and systems management. *IEEE Transactions on Systems, Man and Cybenetics*, 15, 234-243.

Razi, M.A. and Athappilly, K., 2005. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, 29, 65-74.

Rengaswamy, R. and Venkatasubramanian, V., 1995. A syntactic pattern-recognition approach for process monitoring and fault diagnosis. *Engineering Applications of Artificial Intelligence*, 8 (1), 35-51.

Rich, S.H. and Venkatasubramanian, V., 1987. Model-based reasoning in diagnostic expert systems for chemical process plants. *Computers and Chemical Engineering*, 11, 111-122.

Stellenbosch University http://scholar.sun.ac.za

Roberts, S.W., 1959. Control chart tests based on geometric moving averages. *Technometrics*, 1 (3), 239-250.

Rousseeuw, P.J., Ruts, I. and Tukey, J.W., 1999. The Bagplot: A bivariate boxplot. *The American Statistician*, 53 (4), 382-387.

Russell, E., Chiang, L. and Braatz, R., 2000. *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes*. London: Springer-Verlag.

Sathyanarayanamurthy, H. and Chinnam, R.B., 2009. Metamodels for variable importance decomposition with applications to probabilistic engineering design. *Computers & Industrial Engineering*, 57, 996-1007.

Scholz, M., Fraunholz, M. and Selbig, J., 2008. Auto-Associative Models, Nonlinear Principal Component Analysis, Manifolds and Projection Pursuit. *In:* Gorban, A.N., Kegl, B., Wunsch, D.C. and Zinovyev, A., ed. *Principal Manifolds for Data Visualization and Dimension Reduction, Volume 58: Lecture Notes in Computational Science and Engineering*. Berlin: Springer, 202-218.

Schreiber, T., 2000. Measuring information transfer. *Physical Review Letters*, 85, 461-464.

Selltiz, C., Wrightsman, L.S. and Cook, S.W., 1959. *Research Methods in Social Relations*. New York: Holt, Rinehart and Winston.

Shewhart, W.A., 1931. *The economic Control of Quality of Manufactured Product.* New York: Van Nostrand-Reinhold.

Slama, C.F., 1991. *Multivariate statistical analysis of data from an industrial fluidized catalytic process using PCA and PLS.* Thesis (MSc). McMaster University.

Sowa, J.F., 1999. *Processes and Causality* [online]. Available at: http://www.jfsowa.com/ontology/causal.htm [Accessed November 20, 2012].

Steinberg, D. And Colla, P., 1995. *CART: Tree-Structured Nonparametric Data Analysis* [online]. Available at: http://www.salford-systems.com/doc/CART_Main.pdf [Accessed November 20, 2012].

Stockmann, M., Haber, R. and Schmitz, U., 2012. Source identification of plant-wide faults based on k nearest neighbor time delay estimation. *Journal of Process Control*, 22, 583-598.

Sundarraman, A., and Srinivasan, R., 2003. Monitoring transitions in chemical plants using enhanced trend analysis. *Computers and Chemical Engineering*, 27, 1455-1472.

Stellenbosch University http://scholar.sun.ac.za

Sutton, C.D., 2005. Classification and Regression Trees, Bagging, and Boosting. *In:* Rao, C.R., Weqman, E.J. and Solka, J.L., ed. *Handbook of Statistics, Volume 24: Data Mining and Data Visualization*. Amsterdam: Elsevier B.V., 303-329.

Tarifa, E. and Scenna, N., 1997. Fault diagnosis, directed graphs, and fuzzy logic. *Computers and Chemical Engineering*, 21, S649-654.

Taylor, W.A., 2000. *Change-point analysis: A powerful new tool for detecting changes* [online]. Available at: http://www.variation.com/cpa/tech/changepoint.html [Accessed November 20, 2012].

Taylor, A.L., Tait, S.P., Porter, M.A., Perry, M.J. and Nicolson, R.W., 2002. Automatic breakpoint detection for retrospective cumulative sum charts. *Pharmaceutical Statistics*, 1, 25-34.

Teppola, P., Mujunen, S-P., Minkkinen, P., Puijola, T. and Pursiheimo, P., 1998. Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine's wet end. *Chemometrics and Intelligent Laboratory Systems*, 44, 307-317.

Thambirajah, J., Benabbas, L., Bauer, M. and Thornhill, N.F., 2009. Cause-and-effect analysis in chemical processes utilizing XML, plant connectivity and quantitative process history. *Computers and Chemical Engineering*, 33, 503-512.

Thornhill, N. and Horch, A., 2007. Advances and new directions in plant-wide disturbance detection and diagnosis. *Control Engineering Practice*, 15 (30), 1196-1206.

Trahar, W.J. and Warren, L.J., 1976. The floatability of very fine particles – a review. *International Journal of Mineral Processing*, 3, 103.

Tsung, F., 2000. Statistical monitoring and diagnosis of automatic controlled processes using dynamic PCA. *International Journal of Production Research*, 38 (3), 625-637.

Van Lith, P.F., Betlem, B.H.L. and Roffel, B., 2003. Combining prior knowledge with data driven modelling of a batch distillation column including start-up. Computers and Chemical Engineering, 27, 1021-1030.

Van sprang, E.N.M., Ramaker, H.-J., Westerhuis, J.A., Smilde, A.K. and Wienke, D., 2005. Statistical batch process monitoring using gray models. *AIChE Journal*, 51 (3), 931-945.

Vendam, H. and Venkatasubramanian, V., 1997. Signed digraph based multiple fault diagnosis. *Computers in Chemical Engineering*, 21, S655-S660.

Stellenbosch University http://scholar.sun.ac.za

Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S.N., 2003a. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers and Chemical Engineering*, 27, 293-311.

Venkatasubramanian, V., Rengaswamy, R. and Kavuri, S.N., 2003b. A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies. *Computers and Chemical Engineering*, 27, 313-326.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N. and Yin, K., 2003c. A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers and Chemical Engineering*, 27, 327-346.

Verikas, A., Gelzinis, A. and Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44, 330-349.

Villez, K., Rosén, C., Anctil, F., Duchesne, C. and Vanrolleghem, P.A., 2013a. Qualitative Representation of Trends (QRT): Extended method for identification of consecutive inflection points. *Computers and Chemical Engineering*, 48, 187-199.

Villez, K., Venkatasubramanian, V. and Rengaswamy, R., 2013b. Generalized shape constrained spline fitting for qualitative analysis of trends. *Computers and Chemical Engineering*, 58, 116-134.

Wachs, A. and Lewin, D.R., 1999. Improved PCA methods for process disturbance and failure identification. *AIChE Journal*, 45 (8), 1688-1700.

Wan, Y., Yang, F., Lv, N., Xu, H., Ye, H., Li, W., Xu, P., Song, L. and Usadi, A.K., 2013. Statistical root cause analysis of novel faults based on digraph models. *Chemical Engineering Research and Design*, 91, 87-99.

Wang, Y., Cao, F. and Yuan, Y., 2011. *Neurocomputing*, 74, 2483-2490.

Wang, Y., Li, Q., Chang, M., Chen, H. and Zang, G, 2012. Research on Fault Diagnosis Expert System Based on the Neural Network and the Fault Tree Technology. *Procedia Engineering*, 31, 1206-1210.

Westerhuis, J.A. and Coenegracht, P.M.J., 1997. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial east squares. *Journal of Chemometrics*, 11, 379-392.

Westerhuis, J.A., Gurden, S.P. and Smilde, A.K., 2000. Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 51, 95-114.

Stellenbosch University http://scholar.sun.ac.za

Westerhuis, J.A., Kourti, T. and MacGregor, J.F., 1998. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12, 301-321.

Wiener, N., 1956. *The theory of prediction*. New York: McGraw-Hill.

Williams, G.P., 1999. *Chaos Theory Tamed*. Great Britain: Joseph Henry Press.

Wise, B.M. and Gallagher, N.B., 1996. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6 (6), 329-348.

Wold, S., Kettaneh, N. and Tjessem, K., 1996. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, 10, 463-482.

Woodward, R.H. and Goldsmith, P.L., 1964. *Cumulative sum techniques*. Edinburg: Oliver & Boyd for ICI.

Yamanishi, K. And Takeuchi, J., 2002. A unifying framework for detecting outliers and change points from non-stationary time series data. *In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Canada 23-26 July 2002. New York: ACM, 676-681.

Yang, F. and Xiao, D., 2012. Progress in Root Cause and Fault Propagation Analysis of Large-Scale Industrial Processes. *Journal of Control Science and Engineering*, 2012, 1-10.

Yu, J., 2012. Local and global principal component analysis for process monitoring. *Journal of Process Control*, 22, 1358-1373.

Zhang, J., Martin, E.B. and Morris, A.J., 1997. Process monitoring using non-linear statistical techniques. *Chemical Engineering Journal*, 67, 181-189.

Zhang, Y.W., Zhou, H. and Qin, S.J., 2010. Decentralized Fault Diagnosis of Large-scale Processes Using Multiblock Kernel Principal Component Analysis. *Acta Automatica Sinica*, 36 (4), 593-597.

Zhou, Y., Hahn, J. And Mannan, M.S., 2006. Process monitoring based on classification tree and discriminant analysis. *Reliability Engineering and System Safety*, 91, 546-555.

Zhou, Z., Zhuang, M., Lu, X., Hu, L. and Xia, G., 2012. Design of a real-time fault diagnosis expert system for the EAST cryoplant. *Fusion Engineering and Design*, 87, 2002-2006.